

Faktorenanalyse und Selektion

Inaugural-Dissertation
zur
Erlangung der Doktorwürde
der
Naturwissenschaftlichen Fakultät der Philipps-Universität
Marburg/Lahn

vorgelegt von

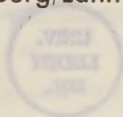
Karl Theo Kalveram
aus Duisburg-Hamborn

Marburg 1968

Fotodruck: Erich Mauersberger, Marburg

Faktorenanalyse und Selektion

Inaugural-Dissertation
zur
Erlangung der Doktorwürde
der
Naturwissenschaftlichen Fakultät der Philipps-Universität
Marburg/Lahn



vorgelegt von

Karl Theo Kalveram
aus Duisburg-Hamborn

Marburg 1968

Fotodruck: Erich Mauersberger, Marburg

Folienanalyse und Selektion

Von der Naturwissenschaftlichen Fakultät der
Philipps-Universität Marburg/Lahn als

Dissertation angenommen am 18.1.1968

Tag der mündlichen Prüfung am 31.1.1968

Berichterstatter: Prof. Dr. F. Merz

Mitberichterstatter: Prof. Dr. A. Peyerimhoff



INHALTSVERZEICHNIS

	Seite
Einleitung.	1
I. Teil: Selektionen und ihre Auswirkungen.	5
A. Einführung in das Problem.	5
1. "Kompensatorische Kovarianz" als Beispiel für einen Selektionseffekt.	5
2. Schwierigkeiten bei der Zusammen- stellung von unausgelesenen Stich- proben.	8
B. Bisher vorliegende theoretische Arbeiten zum Thema Selektion.	11
1. Selektion und Korrelation.	11
2. Selektion und Faktorenanalyse.	12
Die Arbeit von THOMSON & LEDERMANN.	13
Die Arbeit von AHMAVAARA.	15
Die Arbeit von MEREDITH.	18
Sonstige Arbeiten.	21
C. Begründung der eigenen Untersuchungen.	22
II. Teil: Die Darstellung der Selektion im Rahmen des faktorenanalytischen Modells.	27
A. Vereinbarung von Schreibweisen.	27
B. Das lineare Faktorenmodell.	33
1. Grundlegende Definitionen.	33
Grundgleichung und Faktorenmuster.	33
Normierung von Faktoren.	34
Bezeichnungen für spezielle Faktoren.	35
2. Kovarianz und Faktorenmodell.	37
Die Faktorenstruktur.	37
Die Kommunalitäten.	38
Die reduzierte Kovarianzmatrix.	40

Die "reproduzierende" Matrix.	47
Die Existenz von reproduzierenden Matrizen.	49
Die Nichteindeutigkeit der reproduzierenden Matrix.	51
Reproduzierende Matrix und unterlegte Faktoren.	56
Kovarianz, Korrelation und reproduzierende Matrix.	61
Kommunalitäten und reproduzierende Matrix.	63
Der faktorenanalytische Ansatz.	67
C. Die Selektion.	69
1. Selektion und Wahrscheinlichkeitsdichte.	69
Grundlegende Definitionen.	70
Einfache Selektionssätze.	72
Univariate und multivariate Selektion.	75
2. Beispiele für spezielle Arten von Selektionen.	75a
3. Selektion und Kovarianz.	81
Generell zu erwartende Veränderungen.	81
Die Selektionsformeln von LAWLEY.	81
Kompensatorische Kovarianz als spezieller Selektionseffekt.	82
"Ausgeartete" Selektionen.	84
4. Selektion und Faktorenmodell.	86
Konstanz des Faktorenmusters.	86
Veränderung der Kovarianz zwischen Faktoren.	87
Kompensatorische Kovarianz im Faktorenmodell.	88
Selektion und "faktorielle Invarianz".	89
D. Zusammenfassende Darstellung der bisherigen Ergebnisse.	94
E. Diskussion.	97
Zusammenfassung.	109

Anhang.	110
Zahlenbeispiele für Selektionseffekte.	110
1. Beispiel: Trunkation der Summe von zwei unabhängigen zufälligen Variablen.	110
2. Beispiel: Trunkation der Summe von sieben abhängigen zufälligen Variablen.	114
3. Beispiel: Die "Erzeugung" von Faktoren durch Selektion.	122
Literaturverzeichnis.	127



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41554409_0092

Im Jahre 1904 legte SPEARMAN seine Theorie der Intelligenz vor: Einzelne intelligente Verhaltensweisen des Menschen - wie Lesen, Schreiben und Rechnen - wurden als Ausdruck einer "allgemeinen Intelligenz" und einer für die betreffende Tätigkeit spezifischen Fähigkeit angesehen. In der Sprache der Faktorenanalyse lautet diese Theorie: Jeder Intelligenztest einer Testbatterie läßt sich beschreiben mit Hilfe von nur zwei Faktoren: einem Generalfaktor, der der "allgemeinen Intelligenz" und einem spezifischen Faktor, der der auf den Test bezogenen spezifischen Fähigkeit entspricht. Der Faktor der allgemeinen Intelligenz erklärt somit, warum zwei Intelligenztests positiv miteinander korrelieren.

Mit der Zeit wurde offensichtlich, daß ein einziger Faktor häufig nicht in der Lage ist, die Korrelationen zwischen Intelligenztests vollständig zu erklären, sondern daß es hierzu meist mehrerer Faktoren bedarf. Mit der Bereitstellung geeigneter mathematischer Verfahren, die diesem multiplen Ansatz Rechnung trugen (z. B. der "Centroidmethode" THURSTONES), erfuhr die Faktorenanalyse eine bedeutende Weiterentwicklung. Nicht zuletzt diesem Ansatz verdankt die Methode der "multiplen Faktorenanalyse" (THURSTONE 1931) ihre Verbreitung auch in neuerer Zeit.

Faktorenanalytische Techniken werden heute längst nicht mehr allein in der Intelligenzforschung angewendet. Auch im Persönlichkeitsbereich hat die Faktorenanalyse festen Fuß gefaßt: Die Persönlichkeitstheorien z. B. von CATTELL, GUILFORD oder EYSENCK sind - nach Aussagen der Verfasser - faktorenanalytisch gewonnen und fundiert. Andere Gegenstände, die mit Mitteln der Faktorenanalyse untersucht wurden, seien wahllos herausgegriffen und mit jeweils einem Beispiel belegt: Motorische Übungsvorgänge (FLEISHMAN 1954), Soziologie primitiver Gesellschaften (SCHUESSLER & DRIVER 1956), Soziales Verhalten in kleinen Gruppen (BORGOTTA et al. 1956), Moralisches Verhalten von Völkern (CATTELL 1965), Erfassung psychischer Anomalien (ELMORE 1965), Beschreibung von Gesellschaftssystemen (CATTELL

1949), Zusammenhang physiologischer und psychometrischer Merkmale (FAHRENBERG & MYRTEK 1966), Handschriftenanalyse (SEIFERT 1964), Geometrisch-optische Täuschungen (KRISTOF 1964), Aggression (SELG & LISCHKE 1966), Motivation (CATTELL 1963), Angst (BARRAT 1965), Daueraufmerksamkeit (DONALD 1965), Multidimensionale Skalierung (KRISTOF 1963).

Insbesondere fällt auf, daß die Faktorenanalyse zunehmend eingesetzt wird, um Strukturunterschiede zwischen speziellen Gruppen festzustellen. Neben einer kürzlich erschienen theoretischen Arbeit auf diesem Gebiet von MEREDITH (1965) seien auch einige Belege für experimentelle Arbeiten angeführt. So untersuchte LIENERT (1963) die Faktorenstruktur der Intelligenz als Funktion des Neurotizismus, COMREY et al. (1965) verglichen die Persönlichkeitsstruktur von Amerikanern und Italienern, DUNN (1965) beschäftigte sich mit der Stabilität der Faktorenstrukturen von Kindern verschiedenen Alters und Geschlechts hinsichtlich einer neuentwickelten "test anxiety scale", REINERT et al. (1965) untersuchten die Faktorenstruktur von Intelligenztests in Abhängigkeit vom Leistungs- und Begabungsniveau und VUKOVICH (1964) variierte systematisch Probandengruppen und verglich deren Faktorenstrukturen in Bezug auf Begabungstests.

Darüber hinaus ist zu beobachten, daß der klassischen Faktorenanalyse verwandte Verfahren in zunehmendem Maße in Gebrauch kommen, nachdem bisherige faktorenanalytische Modellvorstellungen verallgemeinert und erweitert worden sind; hierher gehören die von TUCKER (1958) und RAO (1948) entwickelte Methode der Zerlegung psychologischer Zeitreihen (z. B. Lernkurven oder Wachstumskurven) in Komponenten, die "three-mode factor analysis" (LEVIN 1965), entwickelt von TUCKER (1964), oder die nicht-metrische Faktorenanalyse KRUSKALS (1964).

Die Häufigkeit faktorenanalytisch ausgerichteter Arbeiten in neuerer Zeit wird von folgenden Befunden belegt:

1. Im zweiten Teil des Bandes 39 der Psychological Abstracts (August bis Dezember 1965) sind rund 7500 Arbeiten auf

allen Teilgebieten der Psychologie erwähnt. Eine Auszählung dieser Arbeiten hat ergeben, daß in 60 Fällen eine Faktorenanalyse empirischer Daten durchgeführt worden ist, während sich 30 Arbeiten mit theoretischen Problemen der Faktorenanalyse beschäftigt haben. Mindestens 90 von 7500 aus aller Welt stammenden Arbeiten eines halben Jahres waren also faktorenanalytisch orientiert, das sind mehr als 1 %.

2. Die Jahrgänge 1953 bis 1966 einschl. von sieben deutschen Zeitschriften wurden auf faktorenanalytisch orientierte Arbeiten hin durchgesehen. Es fanden sich 26 solcher Veröffentlichungen, davon 7 theoretischen Charakters.

3. Von Januar 1964 bis November 1967, also in einem Zeitraum von knapp vier Jahren, wurden allein im Rechenzentrum der Marburger Universität für das Institut für Psychologie mehr als 50 Faktorenanalysen gerechnet.

In letzter Zeit beginnt sich angesichts der Flut von faktorenanalytisch ausgerichteten Arbeiten ein wachsendes Mißbehagen an der Faktorenanalyse deutlich zu artikulieren. Es sind hier nicht nur diejenigen Autoren gemeint, die aus prinzipiellen Erwägungen die Faktorenanalyse ablehnen, sondern auch solche, die sich selbst faktorenanalytischer Verfahren bedienen. So spricht KALLINA (1967, S. 81) vom "Unbehagen in der Faktorenanalyse", FISCHER (1967, S. 136) äußert generell "Zweifel über die Brauchbarkeit der Faktorenanalyse" und ORLIK (1967, S. 87) sieht das "Dilemma der Faktorenanalyse" gar als "Zeichen einer Aufbaukrise in der modernen Psychologie".

Der eigentliche Grund für dieses Unbehagen ist bei den drei Autoren derselbe: Es ist die "erschreckende Unverbindlichkeit" (ORLIK 1967, S. 87), mit der die bisher faktorenanalytisch gewonnenen Ergebnisse nebeneinander stehen. Denn faktorielle Invarianz ist in der Praxis nicht exakt gegeben und kann statistisch nicht überprüft werden. Legt man ähnliche oder gar identische Testbatterien verschiedenen Probandengruppen unter sonst gleichen Umständen vor, so führen Faktorenanalysen der jeweiligen

Daten selten zu Ergebnissen, die wirklich gut übereinstimmen. Hiermit soll nicht gesagt werden, daß ähnliche Resultate in verschiedenen Untersuchungen überhaupt nicht erzielt werden; gleichwohl kann aber auch in solchen Fällen die Übereinstimmung nur nach dem Augenschein beurteilt werden. Sollten aber zwei Ergebnisse, was meistens der Fall ist, in einem oder mehreren Punkten deutlich voneinander abweichen, so kann man über die Ursachen hierfür häufig nur Vermutungen anstellen.

Die Kritik richtet sich nun, so will mir scheinen, bei den letztgenannten Autoren weniger auf die Methode als auf die Theoretiker und Praktiker der Faktorenanalyse: Auf die Theoretiker, weil erstens verschiedene Probleme in der Faktorenanalyse immer noch nicht mathematisch zufriedenstellend gelöst sind und zweitens viel zu wenig Methodenexperimente vorliegen, die den Anwendern das Funktionieren der Methode vor Augen führen könnten. Auf die Praktiker, insofern erstens dem Erfülltsein gewisser Grundvoraussetzungen nur allzuoft keine Beachtung geschenkt wird und zweitens faktorenanalytisch gewonnene Ergebnisse zu selten anderweitig experimentell überprüft werden; allerdings gibt es im Hinblick auf den letztgenannten Punkt in neuerer Zeit einige Ansätze (LIENERT 1964; HEDINGER 1965; HÜRSCH 1967; TAYLOR et al. 1965).

In der vorliegenden Arbeit hingegen wird ein Hauptaugenmerk auf die Methode der Faktorenanalyse selbst gerichtet sein. Es wird u. a. gezeigt, daß sowohl aus logischen als auch aus sachlich begründbaren Erwägungen die Faktorenanalyse - wenigstens in ihrer heute gebräuchlichen Form - nicht als allgemein anwendbare Methode betrachtet werden kann.

I. Teil: Selektionen und ihre Auswirkungen.

A. Einführung in das Problem.

1. Kompensatorische Kovarianz als Beispiel für einen Selektionseffekt.

In der Arbeit "Faktorenanalytische Untersuchungen zur Differenzierungshypothese der Intelligenz: Die Leistungsdifferenzierungshypothese" teilen REINERT et al. (1965) Untertestkorrelationen von Intelligenztests mit, die recht gering, teilweise sogar signifikant negativ ausgefallen sind. Negative Korrelationskoeffizienten bei Intelligenztests sind der Beachtung wert und lassen vermuten, daß sie unerwünschte Effekte einer speziellen Art der Stichprobengewinnung sind.

Eine einfache Überlegung kann erklären, wie es zu solchen niedrigen oder negativen Korrelationskoeffizienten gekommen ist. Man denke sich dazu zwei Intelligenztests an einer größeren Stichprobe A von Probanden angewendet. Nach der Durchführung der Tests werden die beiden Testnoten x_1 und x_2 für jeden Probanden zu einer Gesamtnote zusammengezählt.

Für die weitere Untersuchung nun werden nur solche Probanden berücksichtigt, deren Gesamtnoten bei einem bestimmten Wert, etwa bei 100, liegen. Man bildet also eine kleinere Stichprobe B mit jenen Probanden, deren Testwerte die Bedingung

$$(a) \quad x_1 + x_2 \approx 100$$

erfüllen. Berechnet man nun in der kleineren Stichprobe B den korrelativen Zusammenhang zwischen Test 1 und Test 2, so wird man einen Koeffizienten von nahezu -1 erhalten. In der Stichprobe B nämlich ergänzen sich, wie Bedingung (a) besagt, die beiden Testwerte jeder Versuchsperson zu

100. Wenn also bei einem Probanden das Ergebnis des einen Tests etwa nach oben abweicht, so muß es beim anderen Test im gleichen Maße - sozusagen "kompensatorisch" - nach unten abweichen:

$$x_1 \approx 100 - x_2.$$

Wählt man die kleinere Stichprobe B aus der großen nach dem weniger strengen Kriterium

$$(b) \quad 90 \leq x_1 + x_2 \leq 110$$

aus, wird man natürlich keinen Korrelationskoeffizienten von -1 erwarten; aber es dürfte einleuchten, daß auch hier zwischen den Testergebnissen eine negative "kompensatorische Kovarianz" auftreten muß, da die Variabilität der Summen-Noten eingeschränkt ist.

Was bisher für zwei Tests gesagt wurde, läßt sich leicht verallgemeinern: Bezeichnet man mit x_{i1} , x_{i2} , ..., x_{ik} die Ergebnisse, die Probanden i ($i = 1, 2, \dots, N$) hinsichtlich k Tests erhalten, so wird der beschriebene Effekt immer dann auftreten, wenn mit Hilfe dieses Materials eine neue Stichprobe B wie folgt zusammengestellt wird:

Proband i gehöre zu B, wenn für ihn gilt:

$$(c) \quad a \leq c_1 x_{i1} + c_2 x_{i2} + \dots + c_k x_{ik} \leq b$$

$$(a < b; c_j: \text{Gewichte } (c_j \geq 0)).$$

Für jeden so ausgelesenen Probanden fällt also die - eventuell gewichtete - Summe seiner Testwerte zwischen die beiden fest vorgegebenen Zahlen a und b . Auch hier muß man prinzipiell mit dem Auftreten kompensatorischer Kovarianz rechnen, so daß Korrelationskoeffizienten, die in der unausgelesenen Stichprobe positiv waren, kleiner oder gar negativ werden, während die negativen noch "negativer" werden. Mit wachsendem k wird der Effekt freilich geringer, da sich dann die Varianzeinschränkung der Gesamtnoten auf die einzelne Variable weniger stark auswirkt.

REINERT et al. sind nun tatsächlich genau wie unter (c)

beschrieben vorgegangen, um hinsichtlich der Intelligenz zu relativ homogenen Stichproben zu gelangen.

Bei der eben beschriebenen Art der Stichprobengewinnung handelt es sich um einen speziellen Fall von Selektion, die man auch als "Trunkation" bezeichnet; es ist aber auch denkbar, daß das Auslesekriterium probalistisch ist. Ein solcher Fall liegt z. B. vor, wenn die Auftretenswahrscheinlichkeit der Summenvariable (s. (c)) außerhalb der Grenzen a und b nicht gleich Null, sondern gegenüber dem unselegierten Zustand lediglich herabgesetzt ist. Bei einer solchen probalistischen Selektion sind prinzipiell die gleichen Selektionseffekte zu erwarten wie bei einer Trunkation, wenn auch in abgeschwächtem Ausmaß.

Natürlich wirkt sich eine Selektion nicht nur auf die Kovarianzen zwischen den betroffenen Variablen aus, sondern auch auf deren Varianzen und Mittelwerte. Selektionseffekte hinsichtlich von Mittelwerten sind jedoch für diese Arbeit nicht von Belang, da Mittelwerte keinen Einfluß auf Varianzen und Kovarianzen, die hier allein Gegenstand der Betrachtungen sind, haben.

Kompensatorische Kovarianz ist d e r typische Selektionseffekt, wenn bei einem Merkmal mit komplexer Struktur infolge von Selektionen die Varianz verändert wird. In dem eben beschriebenen Beispiel bestand das komplexe Merkmal aus einer Summe von Einzelmerkmalen; die Varianz dieser Summe wurde durch eine Selektion eingeschränkt. Die Folge war das Auftreten von negativen kompensatorischen Kovarianzen zwischen den Einzelmerkmalen. Würde man die Selektion so durchführen, daß sich die Varianz gegenüber vorher vergrößerte - etwa durch Über-Repräsentation von extremen Ausprägungsgraden der Summe - so wäre die hinzutretende kompensatorische Kovarianz positiv und die Korrelationskoeffizienten zwischen den Einzelmerkmalen würden in positiver Richtung verschoben.

Als Beispiel für ein komplex zusammengesetztes Merkmal nehme man den Intelligenz-Quotienten (IQ) oder einen anderen Kennwert für die Intelligenzleistung: Die meisten

heute gebräuchlichen Intelligenztests bestehen aus mehreren (etwa 10) Untertests. Zur Bestimmung des IQs eines Probanden werden die Untertests getrennt ausgewertet, die so erhaltenen Punktwerte zusammengezählt und die Summe in den IQ-Wert umgerechnet. Ein weiteres Beispiel wäre etwa der Neurotizismus-Wert, den ein Proband in einem Persönlichkeitstest erhält; denn auch dieser Wert setzt sich additiv zusammen aus einer Fülle von Beantwortungen einzelner Test-Items.

Ein ganz anderes Beispiel für ein komplex strukturiertes Merkmal stellt die Beurteilung der "Ähnlichkeit" von bestimmten Reizen dar, wie sie etwa bei der multidimensionalen Skalierung vorkommt. Hier ist das komplexe Merkmal "Ähnlichkeitsurteil" in vielen Fällen allerdings nicht additiv aus der Beachtung einzelner Reizaspekte zusammengesetzt, sondern komplizierterer Natur.

Bei allen diesen komplexen Merkmalen also können Manipulationen, durch die die Varianzen dieser Merkmale in direktem Zugriff verändert werden, zu kompensatorischen Kovarianzen zwischen den konstituierenden Elementar-Merkmalen führen.

Es sei zum Abschluß dieser Ausführungen allerdings darauf hingewiesen, daß kompensatorische Kovarianz nur ein spezieller Selektionseffekt ist und nicht notwendig bei allen Selektionsvorgängen auch die Folge sein muß.

2. Schwierigkeiten bei der Zusammenstellung von unausgelesenen Stichproben.

Die Zusammenstellung einer - bezüglich eines Merkmals oder einer Gruppe von Merkmalen - unausgelesenen Stichprobe ist nun aber in vielen Anwendungsfällen der Korrelations-Statistik mit großen Schwierigkeiten verbunden. Man denke nur daran, welche Mühe es z. B. kostet, will

man hinsichtlich der Intelligenz eine Stichprobe von 11-Jährigen zusammenstellen, die, bezogen auf die Population innerhalb dieser Altersklasse, repräsentativ ist. Schon im mittleren und oberen Intelligenzbereich verteilen sich diese Kinder auf verschiedene Schularten; nämlich auf Volks-, Mittel- und Oberschulen. Kinder im unteren Intelligenzbereich wird man vorwiegend in Sonderschulen zu suchen haben, dort aber befinden sich die Normvarianten zusammen mit asozialen, milieu- und hirngeschädigten Kindern. Die ganz extremen Fälle sind in Heimen untergebracht oder werden in den Familien gehalten. Viele hochbegabte Kinder hingegen laufen Gefahr, völlig übersehen zu werden, insofern sie häufig früher eingeschult werden und/oder Schuljahre überspringen und somit in den altersgerechten Schulklassen nicht anzutreffen sind. Alle diese Kinder aber müssen in einer repräsentativen Stichprobe anteilmäßig richtig vertreten sein.

Bei manchen Problemen ist es völlig unmöglich, zu unausgelesenen Stichproben zu gelangen, etwa wenn der Ausprägungsgrad der interessierenden Merkmale mit einer Größe in Zusammenhang steht, die man als "Mortalitätsrate" oder allgemeiner als Ausfallrate bezeichnen kann.

Als Beispiel hierzu betrachte man die Entwicklung einer Art, bei der die Individuen hinsichtlich verschiedener biologischer Merkmale einem Ausleseprozeß unterliegen. Bei der Anwendung der Korrelationsrechnung auf solche Merkmale ist ebenfalls mit Selektionseffekten zu rechnen, wobei die Individuen, die infolge ihrer speziellen Merkmalskombination ausgeschieden sind, irgendwelchen Messungen überhaupt nicht mehr zugänglich sind.

Analog zur biologischen Auslese sind in der Psychologie diejenigen Fälle, in denen auf Grund der Ergebnisse z. B. eines Berufseignungstestes Schulentlassene verschiedenen Berufen zugeteilt werden. Jede der so gebildeten Gruppen erfährt eine andere Berufsausbildung. Resultate von korrelationsstatistischen Untersuchungen in einer dieser Berufs-

gruppen etwa hinsichtlich des Berufserfolges sind durch Selektionseffekte sicherlich verfälscht. Es ist auch in diesem Falle nicht möglich, die als ungeeignet für den betreffenden Beruf zurückgewiesenen Kandidaten in die Untersuchung einzubeziehen, da sie im Hinblick auf ihre Ausbildung ein anderes Schicksal erfahren.

Wie diese wenigen Beispiele - die mit geringer Mühe um beliebig viele weitere aus der psychologischen Forschung und Praxis vermehrt werden könnten - schon zeigen, sind Stichproben in der Regel ausgelesen. Die Erstellung einer unausgelesenen Stichprobe hingegen erfordert eine sorgfältige Planung und ist daher zumeist mit erheblichen Kosten und - vor allem - mit beträchtlichem Zeitaufwand verbunden. Aus diesen Gründen - und möglicherweise auch aus Unkenntnis - werden bei der praktischen Anwendung der Korrelationsrechnung nur in seltenen Fällen Anstrengungen unternommen, Selektionseffekte zu kontrollieren. Den Beleg für diese Behauptung findet man in den einschlägigen Veröffentlichungen bei der Beschreibung der untersuchten Probandengruppen.

Man möge nicht meinen, Selektionseffekte seien in praktisch belangvollen Fällen so gering, daß man sie ohne weiteres vernachlässigen könne. Wie noch zu besprechende Simulationsexperimente nämlich gezeigt haben, korrelieren zwei ursprünglich unabhängige Variable schon dann zu rund $-0,25$ miteinander, wenn man bezüglich ihrer Summenverteilung - sofern sie normal ist - lediglich 10 % der extremsten Fälle, also links und rechts vom Median je 5 %, ausselektiert. Bei Ausscheidung von 20 % der Extremfälle erniedrigt sich diese Korrelation schon auf rund $-0,4$.

Die bisherigen Ausführungen machen die Bemühungen verschiedener Autoren verständlich, Selektionsvorgänge und ihre Effekte theoretisch besser in den Griff zu bekommen; sie zeigen zugleich auch die Notwendigkeit von Untersuchungen zum Thema "Selektion und Korrelation".

B. Bisher vorliegende theoretische Arbeiten zum
Thema Selektion.

1. Selektion und Korrelation.

Man kann bei einer Selektion die Selektionsvariablen - das sind solche, auf die sich der Auslesevorgang direkt bezieht - in Gedanken unterscheiden von den übrigen Variablen, die nur mittelbar betroffen sind, insofern sie von den Selektionsvariablen stochastisch abhängen. Es erhebt sich dann die Frage: Kann man, wenn der Selektionseffekt hinsichtlich der Selektionsvariablen zahlenmäßig bekannt ist, Mittelwerte, Varianzen und Kovarianzen der übrigen Variablen nicht bereits berechnen?

Der erste, der das Selektionsproblem auf diese Weise angegangen hat, ist Karl PEARSON selbst, auf den bekanntlich der Korrelationskoeffizient in seiner heute gebräuchlichen Form zurückgeht. PEARSON (1903, 1912) löste das Problem für den Fall, daß die Verteilung der Variablen vor Selektion multinormal und die Selektion so beschaffen ist, daß diese Verteilungsform erhalten bleibt. Die von PEARSON angestellten Berechnungen muten jedoch etwas umständlich an und erfordern das Lösen linearer Gleichungssysteme.

Einige Jahre später fand AITKEN (1935) für das PEARSONsche Problem einen eleganteren Lösungsweg, indem er von der momenterzeugenden Funktion ausging und die Matrizen-Schreibweise benutzte. Multinormalität war allerdings auch hier Voraussetzung.

Schließlich folgte eine Arbeit von LAWLEY (1943), in der die Bedingung der Multinormalität fallengelassen wurde und nur die Homoskedastizität und Linearität der Regressionen von Nichtselektionsvariablen auf Selektionsvariable vorausgesetzt wurden. Die eigentliche Bedeutung

der LAWLEYSchen Arbeit besteht in dem Nachweis, daß die spezielle Art der Selektion für die Berechnung des Selektionseffektes nicht bekannt zu sein braucht, gleichgültig, ob es sich z. B. um eine Trunkation oder eine probalistische Auslese handelt. Erforderlich sind lediglich Mittelwerte, Varianzen und Kovarianzen der Selektionsvariablen nach Selektion (Formeln s. S.82).

Eine andere Gruppe von Autoren behandelt das Selektionsproblem in der Weise, daß von der speziellen Art der Selektion auf den Selektionseffekt geschlossen wird. Die Bemühungen in dieser Hinsicht beschränken sich aber bisher sämtlich auf den Spezialfall ein- oder zweiseitiger Trunkation einer einzigen von mehreren Variablen, bei denen die Verteilungsform zudem noch als multinormal vorausgesetzt wird. Die erste Studie dieser Art ist die von CAMPBELL (1945), die sich auf die bivariate Normalverteilung bezieht. 1950 folgen VOTAW, RAFFERTY und DEEMER mit der Behandlung des trivariaten Falles. 1957 schließlich dehnt COHEN diese Überlegungen allgemein auf multinormalverteilte Populationen aus. Weitergehende Verallgemeinerungen liegen anscheinend bisher nicht vor.

Zu erwähnen in diesem Zusammenhang wären eventuell noch BIRNBAUM (1950) sowie BIRNBAUM, PAULSON und ANDREWS (1950). Diese Arbeiten bringen zwar keine neuen Seiten am Selektionsproblem zum Vorschein, zeigen aber, auf welche Weise man Selektionsformeln praktisch bei der Berechnung von Mittelwerten, Varianzen und Kovarianzen der Population nach Selektionen einsetzen kann.

2. Selektion und Faktorenanalyse.

Für die vorliegende Arbeit von größerer Wichtigkeit sind die Autoren, die sich mit den Auswirkungen von Selektionen

auf die Ergebnisse von Faktorenanalysen theoretisch beschäftigt haben.

Gemessen an der großen Zahl von Publikationen über die Faktorenanalyse - HARMAN (1960) zählt in seiner Bibliographie bis 1960 allein 411 Titel auf - gibt es über Selektion und Faktorenanalyse nur sehr wenige Arbeiten. Unter diesen wiederum befinden sich nur drei von Rang: Die Arbeiten von THOMSON & LEDERMANN (1939), AHMAVAARA (1954) und die anscheinend zuletzt auf diesem Gebiet erschienene zusammenfassende Darstellung von MEREDITH (1964). Diese drei Arbeiten seien zunächst besprochen.

Die Arbeit von THOMSON & LEDERMANN.

THOMSON & LEDERMANN (1939) gehen aus von n Tests in standardisierter Form, die sie in p Selektionstests und q Nicht-Selektionstests unterteilen ($n = p + q$) und denen sie folgende Faktorenstruktur unterlegen:

$$(a) \quad \begin{aligned} X &= A_{pr}G + B_{pp}S \\ Y &= A_{qr}G + C_{qq}T. \end{aligned}$$

Hierbei sind mit X die Selektionstests und mit Y die übrigen Tests abgekürzt. G bezeichnet einen Vektor von r Faktoren, die beiden Gruppen von Tests gemeinsam sind, während S und T die den Tests zugehörigen spezifischen Faktoren repräsentieren. A_{pr} , A_{qr} , B_{pp} , C_{qq} bedeuten Matrizen von Ladungen auf diesen Faktoren. Sämtliche Faktoren sind als untereinander stochastisch unabhängig zu denken und sollen die Varianz 1 besitzen.

Bei einer Selektion hinsichtlich der ersten p Tests - so THOMSON & LEDERMANN - treten zu den r gemeinsamen Faktoren noch 1 Gruppenfaktoren hinzu, dies gilt allerdings nur bei den p Selektionstests, nicht bei den übrigen q Tests. Dieses Faktum demonstrieren die Autoren zunächst an einem fiktiven Zahlenbeispiel mit $p = 2$ und untermauern

dieses Ergebnis dann anschließend auch theoretisch. Den Beweisen, für die übrigens LEDERMANN allein verantwortlich zeichnet, liegt folgender Gedankengang zugrunde: Infolge der Selektion werden die r gemeinsamen Faktoren und die ersten p spezifischen Faktoren abhängig, ferner treten auch Korrelationen zwischen gemeinsamen und spezifischen Faktoren auf, während sich hinsichtlich der letzten q spezifischen Faktoren nichts ändert. Mit Hilfe der AITKENschen Selektionsformeln (AITKEN 1935) lassen sich dann die Abhängigkeiten zwischen den Faktoren zahlenmäßig bestimmen und in ihren Auswirkungen verfolgen. Durch eine Reihe von linearen Operationen, die auf die infolge der Selektion abhängigen Faktoren angewendet werden, gelingt es LEDERMANN dann, diese sukzessiv durch andere, aber unabhängige Faktoren zu ersetzen, wobei die ersten p Tests, die Selektionstests, nun durch $r + 1$ gemeinsame Faktoren beschrieben werden, während die letzten $2q$ Tests nach wie vor mit r gemeinsamen Faktoren auskommen. Als obere Schranke für l gibt LEDERMANN den Ausdruck $(2p + 1 - \sqrt{8p + 1})/2$ an.

Obwohl die Autoren nicht ausdrücklich darauf hinweisen, legen sie doch den Schluß nahe, daß man bei Ausschluß der Selektionstests von der Faktorenanalyse zumindest zu der gleichen Anzahl von gemeinsamen Faktoren gelangt, gleichgültig, ob Selektionen nun vorliegen oder nicht.

Abgesehen davon, daß der Rechengang an manchen Stellen reichlich undurchsichtig ist, weisen die Überlegungen LEDERMANNS zwei entscheidende Mängel auf:

a. Daß man abhängige zufällige Variable ersetzen kann durch lineare Kombinationen von unabhängigen zufälligen Variablen, sei unbestritten. Hierauf kommt es jedoch bei dem Selektionsproblem im Rahmen einer Faktorenanalyse nicht an. Ausgangspunkt einer Faktorenanalyse nämlich ist diejenige Korrelationsmatrix der Tests, bei der der Varianzanteil der spezifischen Faktoren durch Subtraktion entfernt worden ist (die sog. "reduzierte" Korrelationsmatrix);

von dieser müßte mindestens gezeigt werden, daß sie sich nach Selektion überhaupt noch faktorisieren läßt. Das aber hat LEDERMANN unterlassen.

b. In den Gleichungen (a) bezeichnet G die gemeinsamen Faktoren, bezogen auf den Rahmen aller $n = p + q$ Tests. Unterteilt man aber diese n Tests in zwei kleinere Gruppen, so kann es sehr wohl vorkommen, daß, bezogen auf eine der so gewonnenen Gruppierungen, aus vorher gemeinsamen Faktoren jetzt spezifische werden. Die Zahl der gemeinsamen Faktoren in der Gruppe der Nicht-Selektionstests kann also schon von vornherein kleiner als r sein. Dieser Umstand ist von LEDERMANN ebenfalls übersehen worden.

Die Arbeit von AHMAVAARA.

Wie THOMSON & LEDERMANN teilt AHMAVAARA (1954) die n beobachteten Testvariablen auf in p Selektionstests und q Nicht-Selektionstests ($n = p + q$) und greift dann ebenfalls auf die AITKENschen Selektionsformeln zurück.

Seine Darstellungsweise ist vorwiegend geometrisch orientiert, da er Tests und Faktoren als Vektoren in entsprechend dimensionierten euklidischen Räumen auffaßt. Seine Ergebnisse gehen aber weit über die Resultate von THOMSON & LEDERMANN hinaus. Sie lauten wie folgt:

1. Selektionen, sofern sie Korrelationen und Varianzen betreffen, sind linearen Transformationen im Selektionsraum (das ist nach AHMAVAARA der Raum, der von den p Selektionstests aufgespannt wird) äquivalent.

2. Eine p -variate Selektion (damit ist eine Selektion bezüglich p Selektionstests gemeint) ist stets rückführbar auf p sukzessiv ausgeführte univariate Selektionen.

3. Bei univariater Selektion wird die Zahl der gemeinsamen Faktoren nicht vermehrt.

4. Bei p -variater Selektion erhöht sich die Zahl der gemeinsamen Faktoren bei den Selektionstests um nicht

mehr als p , während sie bei den übrigen q Tests nicht zunimmt.

5. Drückt man die Tests nach Selektion aus als lineare Kombinationen derselben gemeinsamen Faktoren wie vor der Selektion, so sind die Spalten der neuen Ladungsmatrix proportional den entsprechenden Spalten der ursprünglichen Ladungsmatrix, sofern nicht Korrelationskoeffizienten, sondern Kovarianzen analysiert wurden und die Selektionstests von dieser Analyse ausgeschlossen worden sind.

6. Wenn bei Ausschluß der Selektionstests von der Analyse die verbleibenden Tests eine Einfachstruktur aufweisen, so ist diese Einfachstruktur invariant gegenüber Selektionen.

Sollten diese Feststellungen AHMAVAARAs tatsächlich allgemein zutreffen, so wäre das Selektionsproblem für die Faktorenanalyse als gelöst zu betrachten und gleichzeitig erwiesen, daß es unerheblich ist, ob man selektiertes oder unselektiertes Material einer Analyse unterwirft, da ja die Ergebnisse im wesentlichen invariant wären.

Bei kritischer Betrachtung stellt sich jedoch die Arbeit AHMAVAARAs als mathematisch unbrauchbar heraus, da sie schwerwiegende Mängel und Fehler enthält:

1. Schon die erste Behauptung (bei AHMAVAARA: selektion theorem II), auf der alle weiteren Überlegungen fußen, beruht auf einem Fehlschluß: AHMAVAARA betrachtet lineare Transformationen im Selektionsraum und berechnet deren Auswirkungen auf die Koordinaten der n Testvektoren. Dabei stellt sich heraus, daß die Transformationsgleichungen für die Varianzen und Kovarianzen der Tests dasselbe Aussehen haben wie die von AITKEN mitgeteilten Selektionsformeln. Hieraus folgert AHMAVAARA offenbar erstens, daß jeder linearen Transformation im Selektionsraum eine entsprechend durchgeführte Selektion entspricht und zweitens auch gleich die Umkehrung, daß jede Selektion hinsichtlich

der p Selektionstests sich als lineare Transformation im Selektionsraum darstellen läßt. Dergleichen Schlußweisen sind jedoch in der Mathematik nicht gebräuchlich.

2. Der Ansatz AHMAVAARAs, das Selektionsproblem auf diese Weise geometrisch zu behandeln, erweist sich zudem noch prinzipiell als undurchführbar: AHMAVAARA bezieht die p Selektionstests vor und nach Selektion auf dasselbe Koordinatensystem, bestehend aus p rechtwinkligen Koordinatenachsen. Diesen Koordinatenachsen entsprechen, statistisch gesehen, unabhängige Faktoren $F_1, F_2, \dots, F_p$, mit deren Hilfe also die Selektionstests $X_1, X_2, \dots, X_p$ vor der Selektion wie folgt darstellbar sein sollen:

$$X_i = \sum_{j=1}^p c_{ij} F_j \quad (i=1, 2, \dots, p) \text{ (s. auch S. 33).}$$

Da diese Gleichungen einen funktionellen Zusammenhang zwischen den Selektionstests und den Faktoren ausdrücken, müssen sie in genau derselben Form auch nach erfolgter Selektion gelten; denn eine Selektion betrifft die Auftretenswahrscheinlichkeit von Größen, nicht etwa die Größen selbst (s. S. 36). Da nun die Unabhängigkeit der Faktoren $F_1, \dots, F_p$ durch die Selektion nicht gestört werden soll - im geometrischen Bilde bleiben ja die entsprechenden Koordinatenachsen rechtwinklig -, können sich nur die Varianzen dieser Faktoren geändert haben. Somit folgt hinsichtlich der Varianzen und Kovarianzen der p Selektionstests nach Selektion:

$$\sigma_{X_i}^2 = \sum_{j=1}^p c_{ij}^2 \sigma_{F_j}^2 \quad \text{und} \quad \sigma_{X_i X_k} = \sum_{j=1}^p c_{ij} c_{kj} \sigma_{F_j}^2$$

$$i = 1, 2, \dots, p; \quad k = i + 1, \dots, p.$$

Dieses sind insgesamt $(p^2 + p)/2$ Gleichungen für die p unbekannten neuen Faktorenvarianzen $\sigma_{F_j}^2$. Dieses Gleichungs-

system ist aber schon für $p \gg 2$ i. a. überhaupt nicht mehr lösbar, da mehr Gleichungen als Unbekannte vorliegen. Aber selbst im Falle der Lösbarkeit müßte immer noch nachgewiesen werden, daß der Lösungsvektor nur nicht-negative Komponenten hat, da Varianzen nicht kleiner als Null sein können.

Nebenbei bemerkt sind die eben gemachten Ausführungen ein Beleg dafür, zu welchen Fehlschlüssen man kommen kann, wenn man einen Sachverhalt geometrisch veranschaulicht und sodann das Anschauungsmodell für die Sache selbst hält.

3. Wenn sich bei univariaten Selektionen die Zahl der gemeinsamen Faktoren nicht erhöht und wenn eine p -variante Selektion stets in p aufeinanderfolgende univariate Selektionen aufgelöst werden kann, so darf sich auch bei p -variaten Selektionen die Zahl der gemeinsamen Faktoren nicht erhöhen. AHMAVAARA behauptet aber das Gegenteil.

4. Die Einwände, die vorhin bei der Besprechung der Arbeit von THOMSON & LEDERMANN gemacht worden sind, treffen in gleicher Weise auch bei AHMAVAARA zu. Insbesondere also bietet der Ausschluß der Selektionstests von der Analyse keine Gewähr dafür, daß die Zahl der gemeinsamen Faktoren, bezogen auf die Nichtselektionstests, auch erhalten bleibt.

Die Arbeit von MEREDITH.

MEREDITH geht in seinen Überlegungen von LAWLEYS Selektionstheorem (LAWLEY 1943) aus, das zwar in der Formulierung dem AITKENschen äquivalent ist, aber mit weniger Voraussetzungen auskommt. Wichtigstes Ergebnis bei MEREDITH ist der Nachweis, daß eine Einfachstruktur auch bei Vorliegen von Selektionen invariant bleibt, während dies für andere Ladungsmatrizen als Einfachstrukturen nicht gilt.

Voraussetzungen bei MEREDITH sind Homoskedastizität und Linearität der Regressionen der gemeinsamen Faktoren

auf die Selektionsvariablen, ferner die Unabhängigkeit der Selektionsvariablen von den spezifischen Faktoren. Sowohl gemeinsame als auch spezifische Faktoren sind hierbei auf die zu analysierenden Nicht-Selektionstests bezogen. Diese Bedingungen implizieren Homoskedastizität und Linearität der Regression der (mehrdimensionalen) Nicht-Selektionsvariablen X auf die Selektionsvariable Y und sollen darüber hinaus nach MEREDITH eine direkte Selektion hinsichtlich der Nicht-Selektionsvariable X ausschließen. *)

MEREDITH vermeidet die Schwierigkeit hinsichtlich der Zahl der gemeinsamen Faktoren, die bei Unterteilung der in Betracht stehenden Tests in Selektionstests und Nicht-Selektionstests auftritt, indem er nämlich die Selektionstests nicht in den Rahmen der zu analysierenden Tests einbezieht. Dieser Bezugsrahmen wird somit allein von den Nichtselektionstests gebildet.

Aber auch gegen MEREDITHs Überlegungen kann man Einwände erheben:

1. Das Problem des Nachweises, daß die reduzierte Kovarianzmatrix der manifesten Tests nach Selektion noch positiv semi-definit ist und ihren Rang beibehält, ist nicht gelöst, sondern auf höchst einfache Weise umgangen worden: MEREDITH macht ohne Kommentar implizit die Annahme, daß sich bezüglich der Unabhängigkeit der spezifischen Faktoren nichts ändert (s. S. 179, Übergang von Gl (5) zu Gl (6)), so daß auch nach Selektion die Kovarianz zwischen den Tests wie vorher allein von den gemeinsamen Faktoren herrührt. Natürlich wäre diese implizit gemachte Annahme unter der Voraussetzung, daß die Selektionsvariablen unabhängig von den spezifischen Faktoren

*) Der diesbezügliche Text lautet bei MEREDITH (MEREDITH, S. 178): "These conditions imply the linearity and homoscedasticity of the regression of X on Y and further preclude the possibility of direct selection on X ."

sind, auch unschwer zu beweisen gewesen; letztere Voraussetzung ist aber - nach Ausweis ihrer Begründung (s. Fußnote auf S. 19) - unter diesem Aspekt von MEREDITH offenbar nicht gesehen worden.

2. Die Matrix der Ladungen der Tests auf den Faktoren beschreibt (s. dazu S. 87) den funktionellen Zusammenhang zwischen den Tests und den Faktoren; dieser aber kann als solcher von Selektionen irgendwelcher Art nicht verändert werden. Infolgedessen muß jede Ladungsmatrix gegenüber Selektionen invariant sein. MEREDITH behauptet aber, diese Invarianz träfe lediglich auf solche speziellen Ladungsmatrizen zu, die man als Einfachstrukturen bezeichnen kann.

3. MEREDITH benutzt die LAWLEYSchen Selektionsformeln nur scheinbar dazu, die Kovarianzmatrizen der Tests vor und nach Selektion aufeinander zu beziehen. Nach einer Folge von Substitutionen nämlich gelangt MEREDITH zu folgender wichtiger Formel, die dann Ausgangspunkt weiterer Betrachtungen ist:

$$(a) \quad V_i = F C_i F' + \Delta ,$$

wobei V_i die Kovarianzmatrix der manifesten Tests, C_i die Kovarianzmatrix der ihnen unterlegten gemeinsamen Faktoren nach Selektion und F die Matrix der Ladungen auf diesen Faktoren bedeuten. Δ bedeutet die als Diagonalmatrix vorausgesetzte und durch Selektion nicht (!) veränderte Matrix der spezifischen Varianzanteile (s. MEREDITH 1964 Gl. (7) bzw. (16)). Setzt man aber voraus, daß die betrachteten Tests lineare Kombinationen von Faktoren sind, so kann man Varianzen und Kovarianzen der Tests stets berechnen mit Hilfe der Varianzen der Faktoren und der Kovarianzen zwischen ihnen (s. Folgerung (4.1) und Zerlegungssatz (5)), so daß man unmittelbar zu obiger Gleichung (a) kommt. Dies gilt unabhängig davon, ob Selektionen irgendwelcher Art vorliegen oder nicht. Offenbar also ist MEREDITH hier ein Zirkelschluß unter-

laufen, was allerdings leicht zu erklären ist: Behauptungen, Beweise und daran anschließende Schlußfolgerungen sind bei MEREDITH (1964) nicht leicht auseinanderzuhalten.

Es ist also durchaus zweifelhaft, ob im Hinblick auf die Faktorenanalyse das LAWLEYSche Selektionstheorem tatsächlich "really an extraordinarily powerful tool" ist, wie MEREDITH (MEREDITH, S. 184) es hinstellt.

Sonstige Arbeiten.

Andere Publikationen auf dem Gebiet "Faktorenanalyse und Selektion" sind, gemessen an den eben besprochenen drei Arbeiten, nur von untergeordneter Bedeutung, weil sie sich entweder auf diese stützen oder aber zu keinen weitergehenden Ergebnissen gelangen. So bezieht sich THURSTONE (1945, 1947) bei der Darstellung von Effekten multivariater Selektionen ganz auf die Ergebnisse von THOMSON & LEDERMANN, beschränkt sich aber ansonsten auf Zahlenbeispiele ("boxproblem") und geometrische Veranschaulichungen.

CATTELL sieht das Problem mehr von der praktischen Seite. Er rechnet zwar auch mit einer "distortion of the factorization by special selection of the population" (CATTELL, 1944, S. 274), ist aber offenbar der Meinung, der Effekt solcher Selektionen in Bezug auf die Faktorenanalyse sei lediglich, daß "factors will be present in different amounts, so that the loadings of a set of variables by factor A in the first matrix will appear proportional reduced in the loadings of the same variables by factor A_1 in the second matrix." (CATTELL, 1944, S. 275-276). Nach Aufstellung des "principles of parallel-proportional profile" (CATTELL, 1944, S. 276) beschäftigt er sich in seinen einschlägigen Abhandlungen dementsprechend im wesentlichen mit der Entwicklung von Verfahren zur Erzielung von solchen ähnlichen "factor patterns" (s. z. B. CATTELL, 1949, 1955).

BURT schließlich, der sich ebenfalls theoretisch mit Selektionseffekten beschäftigt hat, geht von PEARSONS

Selektionsformeln (PEARSON, 1912) aus und kommt im Laufe seiner Überlegungen zu einem "proportionality criterion" (BURT, 1939, S. 67-69); denn, so meint BURT - möglicherweise im Hinblick auf Veröffentlichungen von CATTELL -, es sei "better to prove or assess similarity than to postulate it and rotate accordingly" (BURT, 1964, S. 179). Im übrigen jedoch stimmen seine Ergebnisse, von Feinheiten abgesehen, mit denen THOMSONs und LEDERMANNs überein.

C. Begründung der eigenen Untersuchungen.

Es stellt sich also heraus, daß man gegen bisherige Arbeiten zu dem Thema "Selektion und Faktorenanalyse" zahlreiche Einwendungen machen kann. Die wesentlichen Einwände werden im folgenden noch einmal zusammengefaßt, präzisiert und vervollständigt:

1. Es wird behauptet und auch bewiesen, daß eine Einfachstruktur invariant sei gegenüber Selektionen, während dies für andere "factor patterns" i. a. nicht zutreffe. Diese Aussage ist in jeder Hinsicht ohne Sinn:

Ein "factor pattern", aufgefaßt als Matrix von Ladungen auf gegebenen Faktoren, drückt den funktionalen Zusammenhang zwischen den manifesten Tests und den Faktoren aus. Eine solche Matrix kann per definitionem durch Selektionen nicht betroffen sein. Aus diesem Grunde ist jedes "factor pattern", ob es sich nun um eine Einfachstruktur handelt oder nicht, trivialerweise invariant gegenüber Selektionen.

Faßt man ein "factor pattern" auf als "Faktorenlösung", d. h. als Matrix von Zahlen, mit deren Hilfe man die (reduzierte) Korrelationsmatrix reproduzieren kann, so kann die Frage nach Invarianz mathematisch nicht einmal formuliert werden, da eine solche Lösung nur bis auf lineare

Transformationen (z. B. Rotationen) bestimmt und daher nicht eindeutig ist.

2. Es werden nur solche Selektionsvorgänge betrachtet, bei denen die Selektionsvariablen höchstens von den gemeinsamen Faktoren stochastisch abhängen, nicht jedoch von den spezifischen Faktoren - bezieht man gemeinsame und spezifische Faktoren auf die Gruppe der zu analysierenden Nicht-Selektionstests. Die Einteilung in gemeinsame und spezifische Faktoren ist aber eine operationale, insofern sie von dem Rahmen der gerade in Betracht stehenden Tests abhängt (s. hierzu auch WRIGLEY, 1957, S. 82-83). Ein- und derselbe Faktor kann hinsichtlich einer Testbatterie spezifisch sein, hinsichtlich einer anderen jedoch als gemeinsamer Faktor auftreten. Die Voraussetzung der Unabhängigkeit zwischen spezifischen Faktoren und Selektionsvariablen entbehrt daher jeder rationalen Grundlage.

3. Soweit Selektionstests von der Analyse nicht ausgeschlossen werden, fehlen jegliche Angaben darüber, ob die reduzierte Korrelationsmatrix nach Selektion noch positiv-semi-definit ist, d. h. ob die Faktorisierung überhaupt noch möglich ist.

4. Selektionstests und Nicht-Selektionstests werden in bezug auf die Faktorenanalyse unterschiedlich behandelt, indem z. B. die Selektionstests von der Analyse ausgeschlossen werden. Eine Einteilung in Selektionstests und Nicht-Selektionstests kann aber in mathematisch eindeutiger Weise nicht vorgenommen werden; denn erstens ist es schon in einfachen Fällen willkürlich, ob man diese oder jene Variablen als Selektionsvariable ansehen will und zweitens kann man ein System von Selektionsvariablen um beliebige weitere Variable erweitern, also z. B. auch um Nicht-Selektionstests, wobei die Variablen des erweiterten Systems ebenfalls Selektionsvariable im Sinne der Definition sind (s. dazu auch die Ausführungen auf S. 75). Es ist daher absurd, d i e Selektionsvariablen von

einer Faktorenanalyse ausschließen zu wollen.

Es erscheint deshalb geraten, die Theorie der Selektion in bezug auf die Faktorenanalyse neu zu überarbeiten.

Bei dem hierfür notwendigen Rückgriff auf die Grundvorstellungen der Faktorenanalyse kann man sich jedoch eines gewissen Unbehagens nicht erwehren. Es liegen zwar auf diesem Gebiet eine umfangreiche Begriffsbildung und zahlreiche Material vor, man scheut sich aber, die vorliegenden Konstruktionen als "Theorie" zu bezeichnen, denn dafür sind grundlegende Begriffe nur mit mangelnder Präzision gefaßt und wichtige Fragen offengeblieben, möglicherweise sogar übersehen worden. Diese beiden Behauptungen seien zunächst näher erläutert:

1. Eine Korrelationsmatrix ist bei passender Wahl der Hauptdiagonale eine positiv semi-definite Matrix. Mit solchen Matrizen hat man es in der analytischen Geometrie bei der Behandlung von quadratischen Formen ebenfalls zu tun. Infolgedessen gelten alle Operationen und Identitäten bezüglich solcher Matrizen, etwa die Hauptachsen-transformation und die Darstellbarkeit als Produkt einer Matrix A mit ihrer Transponierten A' , selbstverständlich auch für Matrizen von Korrelationskoeffizienten. Ferner gelten für eine Zerlegungsmatrix A einer (reduzierten oder vollständigen) Korrelationsmatrix natürlich auch alle Invarianzsätze gegenüber linearen Transformationen - etwa die Invarianz des Skalarproduktes bei orthogonalen Transformationen.

Nun ist aber der Angelpunkt der eigentlichen Faktorenanalyse nach Ausweis der einschlägigen Lehrbücher stets die Korrelationsmatrix wegen der Möglichkeit ihrer Zerlegung in ein Matrizenprodukt der Form AA' . Also würde man - mathematisch gesehen - erwarten, daß die Faktorenanalyse, wenigstens ab der Einführung der Korrelationsmatrix, formulierbar ist nur in Begriffen der analytischen Geometrie, also ohne Bezug auf statistische Begriffsbildungen.

Variabilität von Meßwerten aber ist konstituierend für die Faktorenanalyse; es werden nämlich erhaltene Meßwerte als Realisationen zufälliger Veränderlicher aufgefaßt und diese wiederum als lineare Kombinationen von Realisationen anderer zufälliger Veränderlicher, den Faktoren. Dieses ist die grundlegende Modellvorstellung, die an bestimmte Sorten von beobachteten Daten herangetragen wird.

Die Aspekte der analytischen Geometrie und der Statistik sind also in der Faktorenanalyse zusammengefügt. Von einer Theorie dieses Gebietes muß man folglich verlangen können, daß sie klar erkennen läßt, wo und wie diese Zusammenfügung geschieht.

Versucht man aber dem nachzugehen, so findet man beide Aspekte - selbst in dem besten zur Zeit vorliegenden Lehrbuch von HARMANN ("Modern Factor Analysis") - durchgehend miteinander verquickt.

2. In sicherlich engem Zusammenhang mit dem eben Gesagten steht die Nichtbearbeitung von gewissen Existenzfragen in der Faktorenanalyse, von denen hier zwei genannt seien:

a. In welcher Beziehung steht eine Koeffizientenmatrix A , die über ein Produkt der Form AA' die Korrelationsmatrix reproduziert, zu den ursprünglichen Meßdaten? Hat insbesondere ein mit Mitteln der analytischen Geometrie gefundenes Ergebnis stets auch statistisch einen Sinn?

b. In Veröffentlichungen und einschlägigen Lehrbüchern wird ohne nähere Begründung, aber offensichtlich der Einfachheit halber, die Unabhängigkeit der sog. spezifischen Faktoren von den sog. gemeinsamen Faktoren unterstellt. Bei Zulassung abhängiger gemeinsamer Faktoren muß man mit der Möglichkeit rechnen, daß spezifische Faktoren von gemeinsamen auch abhängig sein können, da die Einteilung in gemeinsame und spezifische Faktoren vom Rahmen der betrachteten beobachteten Variablen abhängt (s. dazu auch

die Ausführungen auf S. 23). Ist aber in solchen Fällen die zu analysierende reduzierte Korrelationsmatrix mit Sicherheit noch positiv semi-definit, d. h. läßt sie sich überhaupt noch faktorisieren?

Es soll daher im folgenden als erstes versucht werden, die Grundgedanken der Faktorenanalyse so zu formulieren, daß den hier gebrachten Einwänden Rechnung getragen wird. Danach erst werden die bei den Selektionen entstehenden Probleme in Angriff genommen.

II. Teil: Die Darstellung der Selektion im Rahmen des faktorenanalytischen Modells.

A. Vereinbarung von Schreibweisen.

Im folgenden sollen zunächst vereinfachende Schreibweisen verabredet sowie einige grundlegende Begriffe und Sätze mitgeteilt werden. Zu den Definitionen und Sätzen, die sich auf Matrizen beziehen, sei auf GANTMACHER (1965), zu den statistischen Begriffsbildungen auf SCHMETTERER (1956) verwiesen; selbstverständlich können auch andere einschlägige Lehrbücher herangezogen werden.

Matrizen

Mit A_{np} sei eine Matrix von n Zeilen und p Spalten bezeichnet:

$$A_{np} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ . & . & & . \\ . & . & & . \\ a_{n1} & a_{n2} & \dots & a_{np} \end{pmatrix}$$

A_{nn} , B_{pp} sind demnach quadratische Matrizen.

Mit A'_{kn} soll die zu A'_{nk} transponierte Matrix bezeichnet werden; A'_{kn} hat also k Zeilen und n Spalten.

(A_{np}, B_{nq}) sei eine Matrix, die durch Nebeneinanderstellen der beiden Matrizen A_{np} und B_{nq} entsteht; diese Matrix hat n Zeilen und $p + q$ Spalten.

Man überzeugt sich schnell, daß gilt:

$$(A_{np}, B_{nq})' = \begin{pmatrix} A'_{pn} \\ B'_{qn} \end{pmatrix}$$

Die Transponierung ergibt demnach hier eine Matrix von $p + q$ Zeilen und n Spalten. Mit A_{nn}^{-1} wird in der üblichen Weise die zu A_{nn} inverse Matrix bezeichnet.

Sofern keine Verwechslungen möglich sind, werden die Subskripte, insbesondere bei Beweisen, auch weggelassen; statt A_{np} wird dann einfach A geschrieben.

Matrizen der Form

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

(Einheits- oder Identitätsmatrizen) werden stets mit E abgekürzt, Nullmatrizen mit O (Nullmatrizen enthalten als Elemente nur Nullen).

Mit $D_{A_{nn}}$ wird die Diagonalmatrix von A_{nn} bezeichnet, sie enthält außerhalb der Hauptdiagonale nur Nullen, während die Hauptdiagonale selbst mit der Hauptdiagonalen von A_{nn} übereinstimmt.

Zufällige Variable

Die n -dimensionale zufällige Variable $(X_1, X_2, \dots, X_n)$ werde mit X bezeichnet; X hat also n eindimensionale zufällige Variable als Komponenten. Umordnen der Komponenten von X sei u. U. zugelassen.

Eine zufällige Variable Y heiße i -dimensionale Teilvariable von X , wenn sie aus X durch Weglassen von $n - i$ Komponenten entsteht. In diesem Sinne sind X und z. B. X_1 auch Teilvariable von X .

Mit (X, Y) soll diejenige zufällige Variable bezeichnet werden, die durch Nebeneinanderstellen der Komponenten von X und Y entsteht, also $(X, Y) = (X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_m)$; X und Y sind demnach Teilvariable von (X, Y) .

Eine n -dimensionale zufällige Variable X kann formal als einzeilige Matrix mit n Spalten aufgefaßt werden. In X' sind die n Komponenten von X untereinandergeschrieben zu denken.

Realisationen von zufälligen Variablen

Werte von zufälligen Variablen sollen mit entsprechenden kleinen Buchstaben bezeichnet werden, wobei die obigen Schreibweisen und Definitionen sinngemäß übernommen werden sollen. $x = (x_1, x_2, \dots, x_n)$ bezeichnet also eine Realisation der zufälligen n -dimensionalen Variablen X .

Wahrscheinlichkeitsdichte

Mit $f(x_1, x_2, \dots, x_n)$ werde die n -dimensionale Verteilungsdichte, oder auch kurze Dichte, einer n -dimensionalen zufälligen Variablen X bezeichnet (es sei angenommen, daß für alle in dieser Arbeit betrachteten zufälligen Variablen solche Dichten existieren). Verwendet man die obige Vektorsymbolik auch hier, kann die Dichte vereinfacht durch $f(x)$ ausgedrückt werden. Ebenso schreiben wir für die Dichte von (X, Y) vereinfacht $f(x, y)$ anstelle von $f(x_1, x_2, \dots, x_n, y_1, \dots, y_m)$.

Durch Integrationen erhält man aus der Dichte einer mehrdimensionalen zufälligen Variablen die Dichten aller Teilvariablen niedrigerer Dimensionen. So ist z. B. für gegebenes (X, Y)

$$f(x) = \int_{-\infty}^{+\infty} f(x, y) dy;$$

wobei durch $\int_{-\infty}^{+\infty} f(x, y) dy$ angedeutet werden soll, daß die

Integration über den gesamten Definitionsbereich der Integrationsvariablen y ausgeführt werden soll. Ausführlicher geschrieben würde derselbe Ausdruck wie folgt aussehen:

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(x_1, \dots, x_n, y_1, y_2, \dots, y_m) dy_1 dy_2 \dots dy_m.$$

Man sagt auch: $f(x)$ erhält man aus $f(x,y)$ als Randverteilungsdichte.

Bedingte Dichte

Für zwei zufällige Variable X, Y sei die bedingte Verteilungsdichte $f_x(y)$ wie folgt definiert:

$$(1) \quad f_x(y) =: \begin{cases} f(x,y)/f(x) & \text{für } f(x) \neq 0 \\ 0 & \text{für } f(x) = 0 \end{cases}$$

Man nennt $f(x)$ auch die Dichte von Y unter der Hypothese, daß für X die Realisation x vorliegt.

Aus der Definitionsgleichung (1) ergeben sich die folgenden drei Identitäten:

$$(1.1) \quad f(x,y) = f(x) \cdot f_x(y).$$

$$(1.2) \quad f(x,y,z) = f(x) \cdot f_x(y) \cdot f_{(x,y)}(z)$$

$$(1.3) \quad f_x(y,z) = f_x(y) \cdot f_{(x,y)}(z)$$

Varianz und Kovarianz

Mit $E(Z)$ sei der Erwartungswert einer eindimensionalen zufälligen Variablen Z bezeichnet. Für zwei eindimensionale zufällige Variable X_1, X_2 ist dann die Kovarianz definiert als

$$(2) \quad \sigma_{X_1 X_2} =: E \left((X_1 - E(X_1)) \cdot (X_2 - E(X_2)) \right).$$

Im Falle $X_2 = X_1$ erhält man auf diese Weise die Varianz $\sigma_{X_1}^2$ von X_1 .

Für alle in dieser Arbeit vorkommenden Variablen sei vorausgesetzt, daß die Varianzen und damit auch die Kovarianzen existieren.

Über den Ausdruck (2) ist zwischen zwei Variablen X_1 und X_2 eine Verknüpfungsoperation "o" definiert:

$$X_1 \circ X_2 =: \sigma_{X_1 X_2}.$$

Man kann diese Operation auf mehrdimensionale zufällige Variable erweitern, indem man "o" als Analogon zur Matrizenmultiplikation auffaßt: Für zwei zufällige Variable X (k-dimensional) und Y (q-dimensional) sei definiert:

$$X' \circ Y =: \begin{pmatrix} X_1 \circ Y_1 & X_1 \circ Y_2 & \dots & X_1 \circ Y_q \\ \vdots & \vdots & & \vdots \\ X_k \circ Y_1 & X_k \circ Y_2 & \dots & X_k \circ Y_q \end{pmatrix}$$

$X' \circ Y$ ist demnach vom Typ A_{kq} .

Im Falle $Y = X$ bedeutet $X' \circ X$ die Kovarianz-Matrix von X, wobei in der Hauptdiagonale von $X' \circ X$ die Varianzen der Komponenten von X stehen.

Für diese Operation "o" gelten, wie man sofort ein-
sieht, folgende einfache Regeln:

$$(2.1) \quad X' \circ Y = (Y' \circ X)'$$

$$(2.2) \quad A_{nk}(X' \circ Y) B_{qm} = (A_{nk} X') \circ (Y B_{qm})$$

Positiv definite bzw. semi-definite Matrizen.

Eine Matrix A_{nn} heiße positiv definit (bzw. positiv semi-definit), wenn sie symmetrisch ist (d. h. $A = A'$) und für alle $x = (x_1, \dots, x_n)$ gilt:

$$x A_{nn} x' > 0 \text{ (bzw. } x A_{nn} x' \geq 0 \text{)}$$

(2.3) Kovarianz-Matrizen, die in der Hauptdiagonale die Varianzen der Variablen enthalten, sind stets (mindestens) positiv semi-definit.

Die Haupt-Abschnittsdeterminanten einer Matrix A_{nn} seien wie folgt definiert:

$$D_1 =: a_{11}, D_2 =: \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \dots, D_n =: \begin{vmatrix} A_{nn} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}$$

- (2.4) Eine symmetrische Matrix A_{nn} ist genau dann positiv definit, wenn folgende Ungleichungen erfüllt sind:

$$D_1 > 0, D_2 > 0, \dots, D_n > 0.$$

- (2.5) Ist eine Matrix A_{nn} positiv definit, so sind alle ihre Haupt-Unterdeterminanten positiv. (Man erhält eine solche Haupt-Unterdeterminante, wenn man in A_{nn} Zeilen und Spalten mit derselben Numerierung streicht.)
- (2.6) Ist Q positiv definit (bzw. positiv semi-definit), so sind alle Elemente der Hauptdiagonale von Q größer als Null (bzw. mindestens gleich Null).
- (2.7) Q_{nn} ist genau dann positiv definit (bzw. positiv semi-definit), wenn für alle regulären Matrizen C_{nn} gilt: CQC' ist positiv definit (bzw. positiv semi-definit).

B. Das lineare Faktorenmodell.

1. Grundlegende Definitionen.

Grundgleichung und Faktorenmuster.

Gegeben seien n Zufallsvariable $X_1, \dots, X_n$. Diese denke man sich additiv zusammengesetzt aus anderen Zufallsvariablen. Seien also ferner k Zufallsvariable $F_1, \dots, F_k$ und $n \cdot k$ reelle Zahlen a_{ij} ($i = 1, 2, \dots, n$; $j = 1, 2, \dots, k$) so gegeben, daß gilt:

$$\begin{aligned}
 X_1 &= a_{11}F_1 + a_{12}F_2 + \dots + a_{1k}F_k \\
 X_2 &= a_{21}F_1 + a_{22}F_2 + \dots + a_{2k}F_k \\
 &\vdots \\
 &\vdots \\
 &\vdots \\
 X_n &= a_{n1}F_1 + a_{n2}F_2 + \dots + a_{nk}F_k
 \end{aligned}
 \tag{3.1}$$

Die Variablen $F_1, \dots, F_k$ heißen "Faktoren"; sie sind in Anwendungsfällen der Faktorenanalyse meistens nicht direkt beobachtbar, also hypothetischer Natur, während die Variablen $X_1, \dots, X_n$ beobachtbar sind.

a_{ij} heißt "Ladung der Variable X_i auf dem Faktor F_j ".

In den Gleichungen (3.1) drückt sich der funktionale Zusammenhang aus, der zwischen den (beobachteten) Variablen X_i ($i = 1, 2, \dots, n$) und den Faktoren F_j ($j = 1, 2, \dots, k$) bestehen soll. Diese Gleichungen gelten demnach für alle Werte, deren die Variablen X_i und F_j fähig sind. Sie stellen daher eine abkürzende Schreibweise dar für den gleichartigen Zusammenhang zwischen allen möglichen Realisationen der $n + k$ Zufallsvariablen $X_1, \dots, X_n, F_1, \dots, F_k$. Anders formuliert: man denke sich eine Menge von Individuen (besser: Meßwertträgern), bei der jedem Individuum Werte für die Variablen $X_1, \dots, X_n$ und $F_1, \dots, F_k$

zukommen, also insgesamt $n + k$ Meßwerte. Dann geben die Grundgleichungen (3.1) an, auf welche Weise diese Werte bei dem betrachteten Individuum miteinander verknüpft sind.

Die Gleichungen (3.1) gewinnen an Übersichtlichkeit, wenn man sich der Matrix-Schreibweise bedient:

$$\text{Sei } X = (X_1, \dots, X_n); F = (F_1, \dots, F_k); A_{nk} = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nk} \end{pmatrix}$$

dann lassen sich die obigen Gleichungen zusammenfassen zu

$$(3.2) \quad X' = A_{nk} F'.$$

Die Matrix A_{nk} werde als "Ladungsmatrix" oder auch als "Faktorenmuster" bezeichnet; in der angelsächsischen Literatur lautet eine entsprechende Bezeichnung "factor pattern".

Die Gleichungen (3.1) bzw. (3.2) werden oft auch die "Grundgleichungen" der Faktorenanalyse genannt.

Normierung von Faktoren.

In der Praxis sind meistens die Faktoren der Beobachtung nicht zugänglich; in solchen Fällen besteht dann keine Möglichkeit, Zahlenangaben über die Streuungen der Faktoren zu erhalten. Üblicherweise setzt man daher die Streuungen

σ_j der Faktoren willkürlich gleich 1. Damit aber ein Zusammenhang gemäß (3.1) auch nach dieser Normierung erhalten bleibt, müssen die ursprünglichen Ladungen a_{ij} nun durch die Ladungen $b_{ij} = a_{ij} \cdot \sigma_j$ ($\sigma_j \neq 0$) ersetzt werden.

Bezeichnet man mit A_{nk} die Ladungsmatrix vor und mit B_{nk} die Ladungsmatrix nach erfolgter Normierung, so unterscheidet sich eine Spalte j in der Matrix A_{nk} von der entsprechenden Spalte der Matrix B_{nk} lediglich durch einen (konstanten) Proportionalitätsfaktor, nämlich der Streuung σ_j des betreffenden Faktors j vor der Normierung. Es

gilt also:

$$B_{nk} = A_{nk} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \ddots \sigma_k \end{pmatrix}$$

Solche Matrizen seien, der Terminologie CATTELLs folgend, Parallel-Proportional-Matrizen genannt.

Die Unbeobachtbarkeit der Faktoren bringt es mit sich, daß - genau genommen - die Struktur der beobachteten Variablen $X_1, \dots, X_n$ beschrieben wird durch die Klasse aller zu A_{nk} parallel-proportionalen Matrizen; eine spezielle Matrix, etwa die der Ladungen auf den normierten Faktoren, ist lediglich als ein möglicher Repräsentant dieser Klasse anzusehen.

Ist ein Faktor an einer beobachteten Variablen nicht beteiligt, so hat diese Variable auf dem betreffenden Faktor die Ladung Null (eine "Null-Ladung"). An dieser Stelle müssen also alle Ladungsmatrizen der eben eingeführten Äquivalenzklasse parallel-proportionaler Matrizen eine Null aufweisen.

Eine Einfachstruktur, sofern man sie als Ladungsmatrix mit einer außergewöhnlich großen Zahl von solchen Null-Ladungen auffaßt, ist also unempfindlich gegenüber Normierungsoperationen.

Bezeichnungen für spezielle Faktoren.

Zugrunde gelegt seien n beobachtete Variable $X_1, \dots, X_n$, die sich mit Hilfe von k Faktoren $F_1, \dots, F_k$ und einer Ladungsmatrix A_{nk} entsprechend Gleichung (3.2) zusammensetzen.

Ein Faktor heiße ein "gemeinsamer Faktor" ("common factor"), wenn mindestens zwei verschiedene beobachtete Variable nicht-verschwindende Ladungen auf ihm haben.

Ein Faktor heiße ein - im weiteren Sinne - "spezifischer Faktor" ("specific factor"), wenn genau eine beobachtete Variable eine von Null verschiedene Ladung auf ihm hat, die

Ladungen aller übrigen beobachteten Variablen auf ihm aber gleich Null sind.

Liegen p gemeinsame und q spezifische Faktoren vor, so werden erstere mit $G_1, G_2, \dots, G_p$ und letztere mit $H_1, H_2, \dots, H_q$ bezeichnet, wenn dieser Unterschied besonders hervorgehoben werden soll; hierbei ist natürlich $p + q = k$.

Bei sinngemäßer Abkürzung ergibt sich also

$$F = (G, H).$$

Entsprechend kann man auch die Ladungsmatrix A_{nk} aufteilen in die Matrix A_{np} mit den Ladungen auf den gemeinsamen Faktoren und in die Matrix B_{nq} der Ladungen auf den spezifischen Faktoren, wobei die letztere Matrix in jeder Spalte genau eine von Null verschiedene Eintragung besitzt. Somit kann man schreiben:

$$A_{nk} = (A_{np}, B_{nq}),$$

wobei, eventuell nach zweckmäßiger Umordnung der $X_1, \dots, X_n$, B_{nq} wie folgt aussieht:

$$n - q \left(\begin{array}{cccc} b_{11} & \dots & 0 & \\ \vdots & & \vdots & \\ 0 & \dots & b_{qq} & \\ \hline & & & 0 \end{array} \right) \quad \text{bzw.} \quad \left(\begin{array}{cccc|cccc} b_{11} & \dots & 0 & & b_{1,n+1} & \dots & 0 & \\ \vdots & & \vdots & & \vdots & & \vdots & \\ 0 & \dots & b_{nn} & & 0 & \dots & b_{q,n+q} & \\ \hline & & & & & & & 0 \end{array} \right)$$

Mit der Abkürzung $X =: (X_1, \dots, X_n)$ ergibt sich dann

$$(3.3) \quad X' = A_{np} G' + B_{nq} H'.$$

Bemerkung:

Zuweilen unterteilt man die - im weiteren Sinne - spezifischen Faktoren noch in "Fehlerfaktoren" und "spezifische Faktoren im engeren Sinne" und bezeichnet dann die "lineare" Kombination beider Faktoren als "Einzigartigkeitsfaktor" ("unique factor") der betreffenden beobachteten Variablen; mehr als n solche Einzigartigkeitsfaktoren sind dann nicht möglich. Diese Unterteilung soll

hier jedoch nicht getroffen werden, sondern es sei zugelassen, daß die Zahl q der spezifischen Faktoren (im weiteren Sinne) auch größer als n ausfallen kann.

2. Kovarianz und Faktorenmodell.

Die Faktorenstruktur

(4) Definition der Faktorenstruktur:

Gegeben seien eine n -dimensionale zufällige Variable X , eine k -dimensionale Faktor-Variable F und eine Koeffizientenmatrix A_{nk} .

Sei $V_{kk} =: F' \circ F$ (= Kovarianzmatrix von F).

Voraussetzungen:

1. Der Rang aller n -zeiligen und p -spaltigen ($p \leq n$) Teilmatrizen von A_{nk} sei maximal, d. h. gleich p ;
2. V_{kk} sei positiv definit.

(A_{nk} , F , V_{kk}) heie "Faktorenstruktur" von X wenn gilt:

$$X' = A_{nk} F'.$$

Bemerkung:

Obige Voraussetzungen sollen lediglich verhindern, da in einer Faktorenstruktur berflssiges mitgeteilt wird.

4.1) Folgerung

Sei $Q_{nn} =: X' \circ X$ (= Kovarianzmatrix von X).

Dann gilt: $Q_{nn} = A_{nk} V_{kk} A'_{kn}$.

Der Beweis ist schnell gefhrt: Mit Hilfe der Definition

von "o" in (2.1) und der Rechenregel (2.3) folgt unter Beachtung, daß $X = FA'$ ist (Subskripte werden weggelassen):

$$\begin{aligned} Q &= X' \circ X = (AF') \circ (FA') \\ &= A(F' \circ F)A' = A V A'. \end{aligned}$$

Faktoren, Ladungsmatrix und Kovarianzmatrix der Faktoren enthalten also alle Angaben, um die beobachteten Variablen sowie deren Varianzen und Kovarianzen rekonstruieren zu können; aus diesem Grunde wurde auch die Bezeichnung "Struktur" gewählt.

(4.2) Bemerkung:

$Q_{nn} = X' \circ X$ bezeichnet man zuweilen auch als "vollständige" Kovarianzmatrix von X .

Die Kommunalitäten.

Die "gemeinsame" Varianz einer beobachteten Variablen ist ein wichtiger Begriff in der Faktorenanalyse. Diese Varianz ist zu unterscheiden von demjenigen Anteil an der Varianz einer Variablen, der auf die gemeinsamen Faktoren zurückgeht. Es sei daher definiert:

(4.3) Definition:

Gegeben sei eine n -dimensionale zufällige Variable $X = (X_1, \dots, X_n)$ und eine Faktorenstruktur $((A_{np}, B_{nq}), (G, H), V_{kk})$ von X .

Hierbei sei

$G = (G_1, \dots, G_p)$ ein p -dimensionaler gemeinsamer Faktor und $H = (H_1, \dots, H_q)$ ein q -dimensionaler spezifischer Faktor.

Die Elemente von A_{np} seien mit a_{ij} bezeichnet, die Elemente von B_{nq} mit b_{ij} .

$$\text{Sei: } h_i^2 =: X_i \circ (a_{i1}G_1 + a_{i2}G_2 + \dots + a_{ip}G_p),$$

$$k_i^2 =: (a_{i1}G_1 + \dots + a_{ip}G_p) \circ (a_{i1}G_1 + \dots + a_{ip}G_p).$$

h_i^2 heiße die "Kommunalität" von X_i .

k_i^2 heiße die "Kommensurabilität" von X_i .

h_i^2 ist gleich dem Quadrat der multiplen Korrelation zwischen der manifesten Variablen X_i und den gemeinsamen Faktoren $G_1, \dots, G_p$ der Faktorenstruktur, wenn alle Faktoren und die Variablen X_i die Varianz 1 besitzen. h_i^2 ist als diejenige Varianz aufzufassen, die X_i mit den gemeinsamen Faktoren teilt. Im Falle der Abhängigkeit der spezifischen Faktoren von den gemeinsamen tragen also auch die spezifischen Faktoren zu dieser gemeinsamen Varianz von X_i bei.

k_i^2 hingegen ist der Anteil an der Varianz $\sigma_{X_i}^2$ von X_i , der ausschließlich auf die gemeinsamen Faktoren zurückgeht. Spezifische Faktoren haben auf k_i^2 definitionsgemäß keinen Einfluß.

In der Literatur wird zwischen beiden Varianzarten offenbar nicht unterschieden, sondern beide mit dem Namen Kommunalität belegt. Diese Unterlassung ist jedoch ohne Belang, wenn man die Unabhängigkeit der spezifischen Faktoren untereinander und von den gemeinsamen Faktoren voraussetzt - wie dies in der Faktorenanalyse anscheinend stets getan wird.

Folgerungen:

(4.4) Es gilt $k_i^2 \leq h_i^2$ für $i \leq n$.

(4.5) Im Falle $q \leq n$ gilt $h_i^2 = \sigma_{X_i}^2 - b_{ii}^2 \cdot \sigma_{H_i}^2$

($b_{ii}^2 \cdot \sigma_{H_i}^2$ wird auch als "spezifische" Varianz von X_i bezeichnet).

(4.7) Ist H unabhängig von G und sind alle Komponenten von H untereinander unabhängig, so gilt $k_i^2 = h_i^2$.

Beweise:

Zum Beweis nehme man zunächst den Fall $q \leq n$ an. Dann ist $X_i = a_{i1}G_1 + \dots + a_{ip}G_p + b_{ii}H_i$.

Da für "o" - wie für die gewöhnliche Multiplikation auch - das distributive Gesetz gilt, folgen die Behauptungen unmittelbar. Die Verallgemeinerung auf den Fall $q > n$ bereitet keine weiteren Schwierigkeiten.

Die reduzierte Kovarianzmatrix

Unterteilt man die Faktorvariable F in die Faktoren G und H (z. B. in gemeinsame und spezifische Faktoren), so kann man die Kovarianzmatrix der manifesten Variablen X in mehrere Summanden zerlegen:

(5) Hilfssatz (Zerlegungssatz):

Gegeben seien eine n-dimensionale zufällige Variable X und eine Faktorenstruktur $((A_{np}, B_{nq}), (G, H), V_{kk})$ von X ($k = p + q$).

(5.1) Abkürzungen:

$$Q_{nn} =: X'oX \text{ (Kovarianzmatrix von X)}$$

$$P_{nn} =: A_{np}(G'oG)A'_{pn}$$

$$W_{nn} =: A_{np}(G'oH)B'_{qn} + B'_{nq}(H'oG)A'_{pn}$$

$$S_{nn} =: B_{nq}(H'oH)B'_{qn}$$

Behauptung:

$$Q_{nn} = P_{nn} + W_{nn} + S_{nn}$$

Zum Beweis dieses Satzes braucht lediglich der Matrizenkalkül angewandt zu werden.

Faßt man H als gemeinsamen und G als spezifischen Faktor auf sowie A_{np} (bzw. B_{nq}) als Matrix der Ladungen auf G

(bzw. H), so kann man die drei Anteile, aus denen sich die Kovarianz-Matrix Q_{nn} von X zusammensetzt, wie folgt interpretieren:

P_{nn} enthält die auf die gemeinsamen Faktoren zurückgehenden Anteile an den Varianzen und Kovarianzen der Komponenten von X; die Werte in der Hauptdiagonale sind die Kommensurabilitäten von X.

S_{nn} enthält die auf die spezifischen Faktoren zurückgehenden Anteile.

W_{nn} enthält diejenigen Anteile, die auf den Korrelationen zwischen den gemeinsamen Faktoren einerseits und den spezifischen Faktoren andererseits beruhen. Im Falle daß G unabhängig von H ist, ist W_{nn} eine Nullmatrix.

Obiger Hilfssatz gilt aber auch für beliebige andere Unterteilungen der Faktorvariable F.

Es sei nun

G als (p-dimensionaler) gemeinsamer Faktor und

H als (q-dimensionaler) spezifischer aufgefaßt.

$D_{S_{nn}}$ bezeichne die Diagonalmatrix von S_{nn} (s. S. 28); diese Diagonalmatrix enthält nur noch spezifische Varianz-anteile.

(6) Definition der reduzierten Kovarianzmatrix:

Sei $\bar{Q}_{nn} =: Q_{nn} - D_{S_{nn}}$.

$\bar{Q}_{nn}$ heiße die (um die spezifischen Varianzanteile) "reduzierte" Kovarianz-Matrix von X.

$\bar{Q}_{nn}$ und Q_{nn} unterscheiden sich also nur in den Hauptdiagonalen, während die Kovarianzen außerhalb der Hauptdiagonalen in jedem Falle ungeändert bleiben.

Die Werte in der Hauptdiagonale von $\bar{Q}_{nn}$ sind die in (4.3) definierten Kommunalitäten.

Es sei darauf hingewiesen, daß die reduzierte Kovarianzmatrix nur definiert ist in bezug auf eine vorliegende Faktorenstruktur.

Bemerkung:

Eine andere Möglichkeit der Definition der reduzierten Kovarianzmatrix ist folgende:

$$(6.1) \quad \bar{Q}_{nn} =: Q_{nn} - S_{nn}.$$

Im Falle abhängiger spezifischer Faktoren führen (6) und (6.1) zu unterschiedlichen reduzierten Kovarianzmatrizen; denn die Kovarianzanteile, die auf die Korrelationen der spezifischen Faktoren untereinander zurückgehen, werden aus der vollständigen Kovarianzmatrix Q_{nn} bei letzterer Definition ebenfalls entfernt. Ein Streit aber über die Frage, welche der beiden Definitionen der reduzierten Kovarianzmatrix im Falle abhängiger spezifischer Faktoren zweckmäßiger ist, ist müßig: Wie sich herausstellen wird, sind beide Definitionen dann gleichermaßen unzulässig.

Nimmt man jedoch an, daß die spezifischen Faktoren untereinander stochastisch unabhängig sind, so ist $H'oH$ eine Diagonalmatrix. Da B_{nq} nach Definition der spezifischen Faktoren pro Spalte nur eine von Null verschiedene Eintragung besitzt, ist $S_{nn} = B_{nq} (H'oH) B'_{qn}$ ebenfalls eine Diagonalmatrix, d. h. es gilt

$$D_{S_{nn}} = S_{nn},$$

so daß nun die beiden Definitionen der reduzierten Kovarianzmatrix einander äquivalent sind.

(6.2) Definition der doppelt reduzierten Kovarianzmatrix:

Unter denselben Bezeichnungsweisen wie vorhin unter (6) sei:

$$\bar{\bar{Q}}_{nn} =: A_{np} (G' \circ G) A'_{pn}.$$

$\bar{\bar{Q}}_{nn}$ heiße die "doppelt reduzierte" Kovarianzmatrix von X .

In der doppelt reduzierten Kovarianzmatrix $\bar{Q}_{nn}$ sind also nur solche Varianzen und Kovarianzen zwischen den manifesten Variablen aufgeführt, die allein auf die gemeinsamen Faktoren zurückgehen. Die Werte in der Hauptdiagonale von $\bar{Q}_{nn}$ sind die unter (4.3) definierten Kommensurabilitäten. Es sei darauf hingewiesen, daß $\bar{Q}_{nn}$ identisch ist mit der Matrix P_{nn} , definiert unter (5.1).

Ein Hauptinteresse in der Faktorenanalyse richtet sich nun auf den Rang der reduzierten Kovarianzmatrix. Hierzu zunächst ein vorbereitender

(7) Satz (Rangsatz):

Gegeben sei eine n-dimensionale zufällige Variable X, dazu eine Faktorenstruktur $((A_{np}, B_{nq}), (G, H), V_{kk})$ von X, wobei

G ein p-dimensionaler gemeinsamer Faktor und

H ein q-dimensionaler spezifischer Faktor sein soll.

Sei nun H aufgespalten in die t-dimensionale Variable T ($t \leq n$) und die u-dimensionale Variable U; entsprechend sei B_{nq} unterteilt in C_{nt} und D_{nu} , d. h.:

$$H =: (T, U); B_{nq} =: (C_{nt}, D_{nu}); q = t + u; t \leq n.$$

Voraussetzungen:

1. U sei stochastisch unabhängig von (G, T) und
2. alle u Komponenten von U seien untereinander stochastisch unabhängig.

Unter Benutzung der Abkürzungen (5.1) lauten die Behauptungen:

$$(7.1) \quad \bar{Q}_{nn} = A_{np}(G'OG)A'_{pn} + W_{nn} + S_{nn} - D_{S_{nn}}.$$

$$(7.2) \quad \bar{W}_{nn} =: W_{nn} + S_{nn} - D_{S_{nn}} \text{ hat die Gestalt}$$

$$\begin{pmatrix}
 \overbrace{x \dots x}^t & \overbrace{x \dots x}^{n-t} \\
 \vdots & \vdots \\
 \vdots & \vdots \\
 \vdots & \vdots \\
 x \dots x & x \dots x \\
 x \dots x & 0 \dots 0 \\
 \vdots & \vdots \\
 \vdots & \vdots \\
 \vdots & \vdots \\
 x \dots x & 0 \dots 0
 \end{pmatrix}
 \begin{matrix}
 \left. \vphantom{\begin{matrix} x \dots x \\ \vdots \\ \vdots \\ \vdots \\ x \dots x \\ x \dots x \end{matrix}} \right\} t \\
 \left. \vphantom{\begin{matrix} x \dots x \\ \vdots \\ \vdots \\ \vdots \\ 0 \dots 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \dots 0 \end{matrix}} \right\} n-t
 \end{matrix}$$

$$(7.3) \quad \text{Rang } (A_{np}(G' \circ G)A'_{pn}) = p.$$

$$(7.4) \quad \text{Rang } (\bar{Q}_{nn}) \leq p+2t.$$

Beweis zu (7.1): Behauptung folgt unmittelbar aus der Definition (6) und den Abkürzungen (5.1).

Beweis zu (7.2):

Sei $M =: A(G' \circ H)B'$; dann folgt mit (5.1) sofort

$$(a) \quad W_{nn} = M + M'.$$

Da $H = (T, U)$ ist und $G' \circ U = 0$ gilt (G wurde als unabhängig von U vorausgesetzt), folgt:

$$G' \circ H = G' \circ (T, U) = (G' \circ T, G' \circ U) = (G' \circ T, 0).$$

Da $B = (C, D)$ ist, wird

$$(b) \quad M = A(G' \circ T, 0) \begin{pmatrix} C' \\ D' \end{pmatrix} = A(G' \circ T)C'.$$

Wir denken uns nun die Komponenten von X bereits so geordnet, daß C wie folgt aussieht:

$$c = \begin{pmatrix} b_{11} & & 0 \\ & \ddots & \\ 0 & & \ddots & b_{tt} \\ \hline & & 0 & \end{pmatrix} \left. \vphantom{\begin{pmatrix} b_{11} & & 0 \\ & \ddots & \\ 0 & & \ddots & b_{tt} \\ \hline & & 0 & \end{pmatrix}} \right\} n-t$$

Bezeichnet man noch mit r_{ij} die Kovarianz zwischen den Faktoren G_i und T_j , so erhält man für M ausgeschrieben (s. (b)):

$$M = \begin{pmatrix} a_{11} & \dots & a_{qp} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{np} \end{pmatrix} \begin{pmatrix} r_{11} & \dots & r_{1t} \\ \vdots & & \vdots \\ r_{p1} & \dots & r_{pt} \end{pmatrix} \begin{pmatrix} b_{11} & 0 & \dots & \dots & 0 \\ 0 & b_{22} & \dots & 0 & \dots & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & \dots & \dots & b_{tt} & \underbrace{0 \dots 0}_{n-t} \end{pmatrix}$$

$$= \left\{ \begin{array}{cc} \begin{pmatrix} b_{11} & \sum_{i=1}^p a_{1i} r_{i1} & \dots & b_{tt} & \sum_{i=1}^p a_{1i} r_{it} & 0 \dots 0 \\ \vdots & & & \vdots & & \vdots \\ b_{11} & \sum_{i=1}^p a_{ni} r_{i1} & \dots & b_{tt} & \sum_{i=1}^p a_{ni} r_{it} & 0 \dots 0 \end{pmatrix} & \begin{pmatrix} 0 \dots 0 \\ \vdots & \vdots \\ 0 \dots 0 \end{pmatrix} \end{array} \right\} n$$

$\underbrace{\hspace{10em}}_t \qquad \underbrace{\hspace{10em}}_{n-t}$

Also hat zunächst $W_{nn} = M+M'$ (s.(a)) die unter (7.2) angegebene Gestalt.

Weiterhin hat S wegen der vorausgesetzten Unabhängigkeit der letzten $u = q-t$ spezifischen Faktoren die Gestalt

$$\left(\begin{array}{cc} \overbrace{\begin{pmatrix} x & \dots & x \\ \vdots & & \vdots \\ x & \dots & x \end{pmatrix}}^t & \overbrace{\begin{pmatrix} 0 \\ \vdots & \vdots \\ 0 \end{pmatrix}}^{n-t} \\ \hline \begin{pmatrix} 0 & \dots & 0 \end{pmatrix} & \begin{pmatrix} x & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & x \end{pmatrix} \end{array} \right) \quad ,$$

also muß $S - D_S$ wie folgt aussehen:

$$\left(\begin{array}{cc} \overbrace{\begin{pmatrix} 0x & \dots & x \\ x0x & \dots & x \\ \vdots & & \vdots \\ x & \dots & x \end{pmatrix}}^t & \overbrace{\begin{pmatrix} 0 \\ \vdots & \vdots \\ 0 \end{pmatrix}}^{n-t} \\ \hline \begin{pmatrix} 0 & \dots & 0 \end{pmatrix} & \begin{pmatrix} 0 & \dots & 0 \end{pmatrix} \end{array} \right) \quad \left. \begin{array}{l} \vdots \\ \vdots \\ \vdots \end{array} \right\} t$$

$\left. \begin{array}{l} \vdots \\ \vdots \\ \vdots \end{array} \right\} n-t$

Dann hat aber auch $\bar{W} = W + S + D_S$ die in (7.2) behauptete Gestalt.

Beweis zu (7.3):

Die Kovarianzmatrix der Faktoren läßt sich wie folgt unterteilen:

$$V = \left(\begin{array}{c|c} G' \circ G & G' \circ H \\ \hline H' \circ G & H' \circ H \end{array} \right) \quad (k = p + q).$$

V ist nach Definition (4) der Faktorenstruktur positiv definit. Diese Eigenschaft impliziert, daß alle Hauptabschnittsdeterminanten von V, also auch $|G' \circ G|$, größer als Null sind (s. (2.4)). Damit hat $G' \circ G$ den maximalen Rang p.

Da A gemäß Definition (4) ebenfalls den (maximalen) Rang p hat, besitzt auch das Produkt $A(G' \circ G)A'$ den Rang p, was zu zeigen war.

Beweis zu (7.4):

Durch Umstellen der ersten t Zeilen bringen wir $\bar{W}_{nn}$ auf die folgende Form:

$$W^* =: \left(\begin{array}{ccc|cc} \underbrace{x \dots x}_t & \vdots & \vdots & 0 & 0 \\ \vdots & \ddots & \vdots & & \\ \vdots & \vdots & \vdots & & \\ \hline \underbrace{x \dots x}_t & \vdots & \vdots & 0 & 0 \\ \vdots & \ddots & \vdots & & \\ \vdots & \vdots & \vdots & & \\ \hline \underbrace{x \dots x}_t & \underbrace{x \dots x}_{n-2t} & \underbrace{x \dots x}_t & & \\ \vdots & \vdots & \vdots & & \\ \vdots & \vdots & \vdots & & \\ \vdots & \vdots & \vdots & & \\ \hline \underbrace{x \dots x}_t & \underbrace{x \dots x}_{n-2t} & \underbrace{x \dots x}_t & & \end{array} \right) \begin{array}{l} \left. \begin{array}{c} \vdots \\ \vdots \\ \vdots \end{array} \right\} t \\ \left. \begin{array}{c} \vdots \\ \vdots \\ \vdots \end{array} \right\} n-2t \\ \left. \begin{array}{c} \vdots \\ \vdots \\ \vdots \end{array} \right\} t \end{array}$$

Hierbei ist $n > 2t$ angenommen.

Bei einer solchen Umstellung bleibt der Rang der betroffenen Matrix erhalten.

W^* läßt sich also als verallgemeinerte Dreiecksmatrix mit 3 mal 3 konstituierenden Matrizen als Elementen auffassen. Die Determinante von W^* ist damit gleich dem Produkt der Determinanten der Diagonal-Elemente. Dieses Produkt ist aber gleich Null, da das mittlere Element von W^* eine Nullmatrix ist.

Da man alle $(2t + i)$ -reihigen und -spaltigen Teilmatrizen von $\bar{W}$ ($0 < i \leq n - 2t$) ersichtlich auf die gleiche Weise wie eben in eine verallgemeinerte Dreiecksmatrix zerlegen kann, bei der das mittlere Diagonal-Element die Determinante Null hat, haben auch alle Unterdeterminanten der Ordnung $2t + i$ von W^* den Wert Null, d. h. aber:

$$\text{Rang}(\bar{W}) = \text{Rang}(W^*) \leq 2t.$$

Nun hat nach (7.3) $P =: A(G' \circ G)A'$ den Rang p . Da $\bar{Q} = P + \bar{W}$ ist (s. (7.1)), kann der Rang von $\bar{Q}$ die Zahl $p + 2t$ nicht übersteigen.

Im Falle $n \leq 2t$ gilt diese Aussage trivialerweise, da der Rang von $\bar{Q}_{nn}$ die Zahl n nicht übersteigen kann.

Damit ist Satz (7) bewiesen.

(7.5) Folgerung:

Sind sämtliche spezifischen Faktoren untereinander und von allen gemeinsamen Faktoren unabhängig (d. h. $u = q$ und $t = o$), so gilt:

$$\bar{Q}_{nn} = A_{np} (G' \circ G) A'_{pn} = \bar{Q}_{nn}, \text{Rang}(\bar{Q}_{nn}) = p.$$

Die "reproduzierende" Matrix.

Bei der praktischen Anwendung der Faktorenanalyse stellt sich das Problem, aus einem vorliegenden Satz von Daten

eine Matrix von Ladungen der beobachteten Variablen auf (hypothetischen) Faktoren, also eine Faktorenlösung, zu erstellen. Im allgemeinen wendet man hierbei auf die reduzierte Kovarianzmatrix $\bar{Q}_{nn}$ der beobachteten Variablen ein bestimmtes mathematisches Verfahren - etwa die Hauptachsen-Methode - an und erhält eine Matrix A von Zahlen, wobei das Produkt AA' die Ausgangsmatrix $\bar{Q}_{nn}$ reproduziert. Dieser Sachverhalt sei in etwas verallgemeinerter Form wie folgt festgehalten:

(8) Definition der "reproduzierenden" Matrix:

Gegeben sei eine symmetrische Matrix Q_{nn} und eine Matrix C_{nm} vom Range m (also $m \leq n$).

C_{nm} heiße eine " Q_{nn} reproduzierende Matrix", wenn es eine positiv definite Matrix V_{mm} gibt, so daß gilt:

$$Q_{nn} = C_{nm} V_{mm} C'_{mn}.$$

Es sei dem Leser anheimgestellt, V_{mm} als Kovarianzmatrix von Faktoren $G_1, \dots, G_m$ zu interpretieren und C_{nm} als Matrix von Ladungen auf diesen Faktoren; eine Notwendigkeit hierfür besteht nicht.

Zu einer gegebenen symmetrischen Matrix Q_{nn} gibt es - sofern es überhaupt reproduzierende Matrizen gibt - mindestens eine reproduzierende Matrix C_{nm} , bei der die zugehörige Mittenmatrix V_{mm} eine Diagonalmatrix ist. Einer solchen Reproduktionsmatrix kann man also stochastisch unabhängige Faktoren $G_1, \dots, G_m$ unterlegen.

Als "Faktorenanalyse im engeren Sinne" sei eine mathematische Prozedur bezeichnet, die auf eine reduzierte Kovarianzmatrix (bzw. Korrelationsmatrix) $\bar{Q}$ angewendet wird und als Ergebnis eine erste Matrix C liefert, die $\bar{Q}$ reproduziert. Die Hauptachsenmethode (HOTELLING, 1933), die Centroidmethode (THURSTONE, 1947), das Maximum-Likelihood-Verfahren (LAWLEY, 1940), die "kanonische" Faktorenanalyse (RAO, 1955), die Quadratwurzel- bzw. Diagonal-

Methode (HARMAN, 1954) und das Minres-Verfahren (HARMAN & JONES, 1966) sind Beispiele für solche Prozeduren, die darüber hinaus noch Lösungsmatrizen C liefern, für die $Q = CC'$ gilt. Solchen Reproduktionsmatrizen kann man also, sofern es wünschenswert ist, unabhängige Faktoren der Varianz 1 unterlegen.

Die Existenz von reproduzierenden Matrizen.

Eine vorgegebene symmetrische Matrix Q besitzt nun genau dann (mindestens) eine reproduzierende Matrix, wenn die Ausgangsmatrix Q positiv semi-definit ist. Wenn also Q nicht positiv semi-definit ist, so existiert auch keine reproduzierende Matrix zu Q . Ist Q jedoch eine vollständige Kovarianzmatrix, so ist sie auch positiv semi-definit und besitzt daher eine reproduzierende Matrix.

Bei der praktischen Anwendung der Faktorenanalyse geht man - wie gesagt - nicht von der vollständigen, sondern der reduzierten Kovarianzmatrix (bzw. Korrelationsmatrix) der beobachteten Variablen aus. Hier aber ist die Existenz einer Reproduktionsmatrix nicht mehr unbedingt gegeben:

(9.1) Satz:

Nicht jede reduzierte Kovarianzmatrix ist notwendig auch positiv semi-definit, d. h. im allgemeinen existiert zu einer gegebenen reduzierten Kovarianzmatrix keine Faktorenlösung.

Zum Beweis braucht man nur eine reduzierte Kovarianzmatrix anzugeben, die nicht positiv semi-definit ist:

Betrachtet seien zwei (beobachtbare) zufällige Variable X_1, X_2 mit der Varianz 1. Ihnen soll ein gemeinsamer Faktor G_1 und ein spezifischer Faktor H_1 zugrunde liegen, deren Varianzen ebenfalls gleich 1 sind und deren Kovarianz $-0,5$ beträgt. Die Ladungen von X_1 und X_2 auf G_1 mögen $0,6$ und 1 betragen, die Ladungen auf H_1 $0,8$ und 0 . $X = (X_1, X_2)$ besitzt also folgende Faktorenstruktur:

$$\left\{ \left[\begin{pmatrix} 0,6 \\ 1 \end{pmatrix}, \begin{pmatrix} 0,8 \\ 0 \end{pmatrix} \right], (G_1, H_1), \begin{pmatrix} 1 & -0,5 \\ -0,5 & 1 \end{pmatrix} \right\}.$$

Die vollständige Kovarianzmatrix von X, die in diesem Falle identisch ist mit der vollständigen Korrelationsmatrix, errechnet sich gemäß (4.1) zu:

$$\begin{aligned} Q_{22} &= \begin{pmatrix} 0,6 & 0,8 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -0,5 \\ -0,5 & 1 \end{pmatrix} \begin{pmatrix} 0,6 & 1 \\ 0,8 & 0 \end{pmatrix} = \begin{pmatrix} 0,6 & 0,8 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0,2 & 1 \\ 0,5 & -0,5 \end{pmatrix} \\ &= \begin{pmatrix} 0,52 & 0,2 \\ 0,2 & 1 \end{pmatrix}. \end{aligned}$$

Nach (5.1) wird

$$S_{22} = \begin{pmatrix} 0,8 \\ 0 \end{pmatrix} (1) (0,8 \ 0) = \begin{pmatrix} 0,64 & 0 \\ 0 & 0 \end{pmatrix}.$$

Somit erhält man für die reduzierte Kovarianzmatrix nach (6) (oder (6.1))

$$\bar{Q}_{22} = Q_{22} - D_{S_{22}} = \begin{pmatrix} -0,2 & 0,2 \\ 0,2 & 1 \end{pmatrix}.$$

Nun gilt aber der Satz, daß eine Matrix nur dann positiv semi-definit sein kann, wenn alle ihre Haupt-Diagonalelemente nicht negativ sind. Da jedoch in $\bar{Q}_{22}$ das erste Element der Haupt-Diagonale negativ ist, kann Q_{22} nicht positiv semi-definit sein.

Es sei noch hinzugefügt, daß der Wert der Determinante von $\bar{Q}_{22}$ -0,16 beträgt, d. h. der Rang von $\bar{Q}_{22}$ ist 2, obwohl nur ein gemeinsamer Faktor vorhanden ist.

Unter einigen einschränkenden Bedingungen ist aber die Existenz einer reproduzierenden Matrix einer reduzierten Kovarianzmatrix gesichert:

(9.2) Satz:

Betrachtet seien eine n-dimensionale zufällige Variable X und eine Faktorenstruktur $((A_{np}, B_{nq}), (G, H), V_{kk})$ ($k = p + q$) von X. Sei
G p-dimensionaler gemeinsamer und
H q-dimensionaler spezifischer Faktor.

Voraussetzungen:

1. H sei unabhängig von G und
2. Sämtliche Komponenten von H seien untereinander stochastisch unabhängig.

Sei $\bar{Q}_{nn} =$: reduzierte Kovarianzmatrix von X .

Behauptung:

A_{np} ist eine $\bar{Q}_{nn}$ reproduzierende Matrix, d. h. es gibt mindestens eine Matrix, die $\bar{Q}_{nn}$ reproduziert.

Beweis:

Unter den gemachten Voraussetzungen ist nach (7.5) $\bar{Q} = A(G'oG)A'$. $G'oG$ ist als Teilmatrix der positiv definiten Matrix V ebenfalls positiv definit (s. auch Beweis zu (7.3) auf S.46). Da der Rang von A nach Definition (4) der Faktorenstruktur den (maximalen) Wert p hat, ist A nach Definition (8) also eine Matrix, die $\bar{Q}$ reproduziert.

Die Nichteindeutigkeit der reproduzierenden Matrix.

Bekanntlich gibt es viele Matrizen, die eine gegebene positiv semi-definite Matrix reproduzieren: Hat man erst eine solche Reproduktionsmatrix gefunden, erhält man daraus beliebig viele weitere durch Anwendung von linearen Transformationen (z. B. Rotationen):

10) Transformationssatz:

Betrachtet sei eine symmetrische Matrix Q_{nn} . Sei ferner L_{mm} eine reguläre (Transformations-)Matrix.

Behauptung:

$C_{nm} L_{mm}$ reproduziert Q_{nn} dann und nur dann, wenn Q_{nn} von C_{nm} reproduziert wird.

Beweis: zu "dann":

- (a) C reproduziere Q, d. h. es gibt eine positiv definite Matrix V, so daß nach Definition (8) $Q = C \cdot V \cdot C'$ ist.
- (b) Sei nun $W =: L^{-1} V (L^{-1})'$ und $D =: C L$.

W ist nach (2.7) ebenfalls positiv definit, da mit L auch L^{-1} regulär ist. D hat den (maximalen) Rang m, da C als Reproduktionsmatrix nach Definition (8) und L nach Vorgabe beide den maximalen Rang m haben.

Nun gilt:

$$D W D' = C L (L^{-1} V (L^{-1})') (C L)' = C \underbrace{L L^{-1}}_E V \underbrace{(L')^{-1} L'}_E C' = Q. \quad (a)$$

(b)

Also ist $D = C L$ eine Matrix, die nach Definition (8) Q reproduziert.

Beweis zu "nur dann":

C L reproduziere Q. L^{-1} ist regulär, also reproduziert nach dem gerade Bewiesenen $(C L) L^{-1} = C (L L^{-1}) = C$ ebenfalls Q.

Damit ist Satz (10) bewiesen.

Das Vollständigkeitsproblem.

Setzt man voraus, daß zu einer gegebenen reduzierten Kovarianzmatrix Reproduktionsmatrizen existieren, stellen sich folgende Fragen:

a. Erhält man durch Anwendung von linearen Transformationen auf eine reproduzierende Matrix tatsächlich auch alle reproduzierenden Matrizen?

b. Ist diejenige Ladungsmatrix, die in der Faktorestruktur ursprünglich vorgegeben war, ganz oder wenigstens zum Teil in der Menge der reproduzierenden Matrizen enthalten? D. h. ist man mit Hilfe einer Faktorenanalyse in der Lage, eine tatsächlich vorliegende Merkmalsstruktur - im Prinzip wenigstens - wiederzufinden?

Diesen Fragen sei jetzt in zwei Schritten nachgegangen.

(11) Vollständigkeitssatz:

Betrachtet seien eine symmetrische Matrix Q_{nn} und zwei Matrizen C_{nm} , D_{nm} , die beide Q_{nn} reproduzieren (als Reproduktionsmatrizen von Q_{nn} können C und D keinen unterschiedlichen Rang und damit keine unterschiedlichen Spaltenzahlen haben).

Behauptung:

Es gibt genau eine reguläre lineare Transformation L_{mm} , so daß gilt:

$$D_{nm} = C_{nm} L_{mm}.$$

Diese Transformation sieht wie folgt aus:

$$L_{mm} = (C'_{mn} C_{nm})^{-1} C'_{mn} D_{nm}.$$

Beweis:

Entsprechend der Definition (8) seien positiv definite Matrizen V und W vorgegeben, so daß gilt:

$$(a) \quad C V C' = Q = D W D'$$

$$(b) \text{ Sei: } L =: (C' C)^{-1} C' D \quad \text{und} \quad M =: W D' D$$

Alle an den obigen Produkten beteiligten Matrizen haben den maximalen Rang m . Es existieren daher $(C' C)^{-1}$ und $(D' D)^{-1}$, also auch $(W D' D)^{-1}$. Die Regularität von L ist aus demselben Grunde ebenfalls gesichert.

Es folgt:

$$CL = C'(C'C)^{-1}C'D \underbrace{(MM^{-1})}_E = C(C'C)^{-1} \underbrace{C'D(WD'D)}_{CVC'} (WD'D)^{-1} \quad (b) \quad n. (a)$$

$$= C(C'C)^{-1}C'(CVC')D(WD'D)^{-1}$$

$$= C \underbrace{(C'C)^{-1} (C'C)}_E VC'D (WD'D)^{-1} = (CVC')D (WD'D)^{-1}$$

$$= (DWD')D (WD'D)^{-1} = D \underbrace{(WD'D) (WD'D)^{-1}}_E = D.$$

$$(a)$$

Damit ist der erste Teil der Behauptung bewiesen.

Zum Beweis der Eindeutigkeit betrachte man zwei Matrizen L, K mit $D = CL$ und $D = CK$. Es folgt $CL = CK$. Linksmultiplikation dieser Gleichung mit $(C'C)^{-1}C'$ ergibt $L = K$.

(11.1) Folgerung:

Sind die beiden Matrizen C und D , die die Matrix Q reproduzieren, speziell so beschaffen, daß

$$(a) \quad C C' = Q = D D'$$

gilt, (d. h. ist die zugehörige Beimatrix V in beiden Fällen eine Einheitsmatrix) so ist die Transformation, die C in D überführt, orthogonal.

Zum Beweis ist lediglich noch zu zeigen, daß für die Transformation $L =: (C'C)^{-1} C'D$, für die ja nach (11) $D = CL$ gilt, die Orthogonalitätsbedingung $LL' = E = L'L$ erfüllt ist:

Zunächst gilt

$$L' = D'C((C'C)^{-1})' = D'C((C'C)')^{-1} = D'C(C'C)^{-1}.$$

Damit folgt

$$LL' = (C'C)^{-1} C' \underbrace{D D'}_{C C' \text{ n. (a)}} C(C'C)^{-1} = (C'C)^{-1} (C'C) (C'C) (C'C)^{-1}$$

$$= E = L'L.$$

Hauptachsenmethode, Centroidmethode, Diagonalmethode und andere, in der Faktorenanalyse heute gebräuchliche Verfahren liefern reproduzierende Matrizen, die die unter Folgerung (11.1) beschriebene spezielle Beschaffenheit haben; man kann ihnen also unabhängige Faktoren mit der Varianz 1 unterlegen. Allerdings liefert jedes Verfahren eine andere Reproduktionsmatrix. Der Folgerung (11.1) kann man jedoch entnehmen, daß alle diese Matrizen einander völlig gleichwertig sind, insofern sie sich durch orthogonale Transformationen, also durch Drehungen und Spiegelungen, ineinander überführen lassen.

Bei Vorliegen einer reproduzierenden Matrix werden also nach (11) alle übrigen reproduzierenden Matrizen, die Q reproduzieren, zugänglich mit Hilfe regulärer linearer Transformationen.

Nun kann auch geklärt werden, inwieweit eine ursprünglich vorgegebene Faktorenstruktur durch eine Faktorenanalyse auffindbar ist:

(12) Satz:

Betrachtet seien eine n-dimensionale zufällige Variable X und eine Faktorenstruktur $((A_{np}, B_{nq}), (G, H), V_{kk})$ ($k = p + q$) von X.

Sei G p-dimensionaler gemeinsamer Faktor und
H q-dimensionaler spezifischer Faktor.

Voraussetzungen:

1. H sei stochastisch unabhängig von G und
2. alle q Komponenten von H seien untereinander stochastisch unabhängig.

Dann hat nach (7.5) die reduzierte Kovarianzmatrix $\bar{Q}_{nn}$ von X den Rang p und besitzt auch (nach (9.2)) eine reproduzierende Matrix.

Behauptung:

Zu jeder Matrix C_{np} , die $\bar{Q}_{nn}$ reproduziert, gibt es eine lineare reguläre Transformation L_{pp} , so daß gilt:

$$A_{np} = C_{np} L_{pp} \quad (\text{bzw. } C_{np} = A_{np} L_{pp}^{-1}).$$

Der Beweis ergibt sich mit Hilfe der Sätze (9.2), (11).

Die beiden Voraussetzungen, die Unabhängigkeit der spezifischen Faktoren betreffend, gewährleisten also die (prinzipielle) Auffindbarkeit der Matrix A_{np} der Ladungen der manifesten Variablen auf den gemeinsamen Faktoren $G_1, \dots, G_p$ mit Hilfe einer Faktorenanalyse. Ob diese Unabhängigkeitsvoraussetzungen sachlich gerechtfertigt sind, muß

im Einzelfall entschieden werden.

Ist die Unabhängigkeit der spezifischen Faktoren untereinander und/oder von den gemeinsamen Faktoren nicht gegeben, so ist nicht einmal sicher, daß es überhaupt Matrizen gibt, die die reduzierte Kovarianzmatrix $\bar{Q}_{nn}$ reproduzieren. Wenn das ausnahmsweise doch der Fall sein sollte, ist die - ursprünglich an der Faktorenstruktur beteiligt gewesene - Ladungsmatrix A_{np} nicht notwendig unter den durch eine Faktorenanalyse auffindbaren Reproduktionsmatrizen; denn bei Verletzung der Bedingung der Unabhängigkeit ist im allgemeinen der Rang von $\bar{Q}_{nn}$ (s. Rangsatz (7.4)) größer als p und damit haben auch alle Reproduktionsmatrizen von $\bar{Q}_{nn}$ mehr als p Spalten. Lineare reguläre Transformationen, die solche Matrizen in A_{np} überführen könnten, existieren nicht.

Reproduzierende Matrix und unterlegte Faktoren.

Eine reproduzierende Matrix kann man linearen Transformationen unterwerfen, ohne daß sie ihre Reproduktionseigenschaften verliert (s. Transformationssatz (10)). Hiervon macht man denn auch in der Faktorenanalyse ausgiebig Gebrauch. Die durchgeführten Transformationen - meist handelt es sich um orthogonale - dienen meistens dem Ziel, eine bessere Interpretierbarkeit der Ladungsmatrizen durch Überführung in eine sog. Einfachstruktur zu erreichen. Eine Einfachstruktur ist, mehr schlecht als recht beschrieben, eine Reproduktionsmatrix mit außergewöhnlich vielen Nullen. Kriterien dafür, ob man nach Durchführung einer Transformation als Ergebnis die Matrix der Ladungen auf den Ausgangsfaktoren erhalten oder wenigstens angenähert hat, gibt es bisher nicht.

Es erhebt sich nun die Frage, ob man solcherart transformierte Reproduktionsmatrizen auffassen kann als Teil einer Faktorenstruktur, die von der ursprünglich unterlegten möglicherweise verschieden ist.

Diese Frage sei zunächst präziser gefaßt:

Ausgegangen sei wieder von einer beobachteten zufälligen Variablen $X = (X_1, \dots, X_n)$ mit der Faktorenstruktur $((A_{np}, B_{nq}), (G, H), V_{kk})$ ($k = p + q$), wobei $G =: (G_1, \dots, G_p)$ ein p -dimensionaler gemeinsamer Faktor und $H =: (H_1, \dots, H_q)$ ein q -dimensionaler spezifischer Faktor sein soll.

A_{np} und B_{nq} seien die Ladungsmatrizen bezgl. dieser Faktoren.

Sei ferner $p, q \leq n$.

Entsprechend der Definition der Faktorenstruktur (4) gilt

$$(a) \quad X' = A_{np} G' + B_{nq} H'.$$

Diese Gleichung gibt in abgekürzter Schreibweise den funktionalen Zusammenhang zwischen den $X_1, \dots, X_n$ wieder.

Der Einfachheit halber sei nun angenommen, daß sämtliche Faktoren voneinander unabhängig sind und die Varianz 1 besitzen; V_{kk} ist dann eine Einheitsmatrix E .

Die Variablen $X_1, \dots, X_n$ mögen ebenfalls die Varianz 1 besitzen.

Bezeichnet man mit $\bar{Q}_{nn}$ die reduzierte Kovarianzmatrix von X , so gilt

$$\bar{Q}_{nn} = A_{np} A'_{pn}.$$

Sei nun C_{np} irgendeine Matrix, die $\bar{Q}_{nn}$ reproduziert und die aus A_{np} durch orthogonale Transformation hervorgeht. Es gilt dann ebenfalls

$$\bar{Q}_{nn} = C_{np} C'_{pn}.$$

Dieser Matrix C_{np} seien nun stochastisch unabhängige Faktoren $F_1, F_2, \dots, F_p$ der Varianz 1 unterlegt, d. h. C_{np} werde interpretiert als Matrix von Ladungen auf diesen Faktoren.

Mit der Abkürzung $F =: (F_1, \dots, F_p)$ lautet die oben gestellte Frage in präziser Fassung: Ist auch $((C_{np}, B_{nq}),$

(F,H), E) eine Faktorenstruktur von X, d. h. gilt

$$(b) \quad X' = C_{np} F' + B_{nq} H'$$

ebenfalls? Mit anderen Worten: Ist durch C_{np} und F ein anderer funktionaler Zusammenhang zwischen den Variablen $X_1, \dots, X_n$ als der unter (a) aufgeführte gegeben?

Der Literatur über die Faktorenanalyse ist kein Hinweis darauf zu entnehmen, daß an der Bejahung dieser Frage irgendwelche Zweifel bestehen könnten; im Gegenteil wird das Bestehen auch der Gleichung (b) offenbar fraglos als gegeben angesehen und sogar hin und wieder als Ausgangspunkt für erkenntnistheoretische Spekulationen benutzt (s. z. B. HARMAN, 1960, S. 9 u. S. 21; KALLINA, 1964 *)).

Eine gewisse Gleichwertigkeit der ursprünglichen Faktorenstruktur und der formal daraus abgeleiteten ist natürlich insoweit gegeben, daß sowohl A_{np} als auch C_{np} beide die reduzierte Kovarianzmatrix reproduzieren. Das allein aber reicht noch nicht aus; denn eine Faktorenanalyse soll ja schließlich den (bzw. einen möglichen) funktionalen Zusammenhang zwischen den beobachteten Variablen aufdecken, d. h. in die Lage versetzen, die Daten selbst zu reproduzieren, und zwar über eine Gleichung der Form (a) bzw. (b) - nun ausgeschrieben gedacht für einzelne Realisationen.

In diesem Sinne aber ist die Gleichwertigkeit der beiden Faktorenstrukturen nicht mehr unbedingt gegeben; es kann nämlich sehr wohl vorkommen, daß es zu einer Matrix, die $\bar{Q}_{nn}$ richtig reproduziert, überhaupt keine Faktoren so gibt, daß die Daten selbst reproduziert werden könnten. D. h.: Mit Hilfsmitteln der analytischen Geometrie gewonnene Ergebnisse in der Faktorenanalyse brauchen - statistisch gesehen (bzw. auf die erhaltenen Daten bezogen) - keinen Sinn zu ergeben.

An einem einfachen Beispiel soll das nun demonstriert

*) persönliche Mitteilung.

werden:

Für $n = 3$, $p = 2$, $q = 0$ und $A_{32} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$ lautet

Gleichung (a) ausgeschrieben:

$$(a1) \quad \begin{aligned} X_1 &= G_1 \\ X_2 &= G_2 \\ X_3 &= 1/\sqrt{2}G_1 + 1/\sqrt{2}G_2; \end{aligned}$$

wobei G_1 , G_2 voneinander stochastisch unabhängig sind und die Varianz 1 besitzen.

Zusätzlich sei vorausgesetzt, daß die Dichten von G_1 und G_2 wie folgt aussehen:

$$(c) \quad f(g_1) = \begin{cases} 1/\sqrt{6} & \text{für } -\sqrt{3} \leq g_1 \leq \sqrt{3} \\ 0 & \text{sonst} \end{cases}; \quad f(g_2) \text{ analog.}$$

G_1 und G_2 sind also innerhalb ihres Variationsbereichs $\pm\sqrt{3}$ gleichverteilt.

$$(d) \quad \text{Nach (a1) sind dann } X_1 \text{ und } X_2 \text{ ebenfalls gleichverteilt im Intervall } -\sqrt{3} \dots \sqrt{3}.$$

Die Dichte von X_3 hat die Form eines gleichseitigen Dreiecks:

$$(e) \quad f(x_3) = \begin{cases} (\sqrt{6} + x_3) / 6 & \text{für } -\sqrt{6} \leq x_3 \leq 0 \\ (\sqrt{6} - x_3) / 6 & \text{für } 0 \leq x_3 \leq \sqrt{6} \\ \text{sonst.} & \end{cases} \quad *)$$

Durch Anwendung der orthogonalen Transformation

$$L_{22} =: \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix}, \quad \alpha = 45^\circ,$$

auf A_{32} (Drehung des Koordinatensystems um 45°) erhält man als "neue" Reproduktionsmatrix von $\bar{Q}_{nn}$

$$G_{32} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1 \end{pmatrix}$$

*) Diese Dichte erhält man aus (a1) durch eine sog. "Faltung".

Angenommen nun, es gäbe Faktoren F_1 und F_2 solcherart, daß sich die beobachteten Variablen in folgender Form schreiben lassen:

$$\begin{aligned} X_1 &= 1/\sqrt{2} F_1 + 1/\sqrt{2} F_2 \\ (f) \quad X_2 &= 1/\sqrt{2} F_1 - 1/\sqrt{2} F_2 \\ X_3 &= F_2, \end{aligned}$$

wobei F_1 und F_2 unabhängig sind und beide die Varianz 1 haben.

F_2 variiert dann innerhalb der Grenzen $\pm \sqrt{6}$ wegen der Identität von F_2 mit X_3 (s. (f) (e)). $1/\sqrt{2} F_2$ variiert also innerhalb $\pm \sqrt{3}$. Da ferner die Varianz von F_1 gleich 1 ist, muß $1/\sqrt{2} F_1$ eine von Null verschiedene Variationsbreite besitzen. Bei unabhängigen zufälligen Variablen ist aber die Variationsbreite ihrer Summe gleich der Summe der Variationsbreiten; somit muß $1/\sqrt{2} F_1 + 1/\sqrt{2} F_2$ auch außerhalb $\pm \sqrt{3}$ variieren. Versucht man also, über die angenommene Gleichung (f) die beobachteten Variablen zu rekonstruieren, so kommt man zu Werten für X_1 , die in der Tat gar nicht vorkommen können; denn X_1 variiert nach (d), (c) genau innerhalb der Grenzen $\pm \sqrt{3}$. Schon aus diesem Grunde kann es also keine Faktoren geben, die die Gleichungen (f) erfüllen.

Tatsächlich ergeben sich aus der obigen Annahme (f) noch weitere Widersprüche; denn $1/\sqrt{2} F_1 + 1/\sqrt{2} F_2$ muß eine Gleichverteilung besitzen, eine Forderung, die sich bei unabhängigen Faktoren mit von Null verschiedener Varianz überhaupt nicht erfüllen läßt.

Somit ist erwiesen, daß es zu einer vorliegenden Faktorenlösung nicht notwendig auch solche Faktoren zu geben braucht, daß über eine Gleichung der Form (a) (Grundgleichung) die beobachteten Variablen reproduziert werden.

Kovarianz, Korrelation und reproduzierende Matrix.

Um den Korrelationskoeffizienten zwischen zwei Variablen X_1 und X_2 zu erhalten, muß die Kovarianz $X_1 \circ X_2$ zwischen ihnen durch das Produkt der Variablenstreuungen dividiert werden:

$$r_{12} = \frac{X_1 \circ X_2}{\sigma_{X_1} \cdot \sigma_{X_2}} .$$

Definiert man für eine gegebene zufällige Variable $X = (X_1, \dots, X_n)$

$$D =: \begin{pmatrix} \sigma_{X_1}^{-1} & & 0 \\ & \ddots & \\ 0 & & \sigma_{X_n}^{-1} \end{pmatrix}$$

und bezeichnet man mit Q_{nn} die (vollständige) Kovarianzmatrix von X , so erhält man die vollständige Korrelationsmatrix R_{nn} von X wie folgt:

$$(13.1) \quad R_{nn} = D Q_{nn} D$$

Die reduzierte Korrelationsmatrix $\bar{R}_{nn}$ von X ist dann gegeben durch

$$(13.2) \quad \bar{R}_{nn} = D \bar{Q}_{nn} D,$$

wobei $\bar{Q}_{nn}$ die - hinsichtlich einer Faktorenstruktur von X definierte - reduzierte Kovarianzmatrix bedeutet.

Da D eine Diagonalmatrix ist mit nichtverschwindenden Elementen in der Hauptdiagonale, gilt $\text{Rang}(D) = n$, so daß $\bar{R}_{nn}$ denselben Rang wie $\bar{Q}_{nn}$ besitzt. Ebenso wie $\bar{Q}_{nn}$ ist auch $\bar{R}_{nn}$ symmetrisch.

Zwischen R_{nn} und Q_{nn} (bzw. $\bar{R}_{nn}$ und $\bar{Q}_{nn}$) bestehen nun hinsichtlich der sie jeweils reproduzierenden Matrizen folgende Beziehungen:

(13.3) Satz:

Behauptungen:

1. Ist C_{np} eine Q_{nn} (bzw. $\bar{Q}_{nn}$) reproduzierende Matrix, so wird R_{nn} (bzw. $\bar{R}_{nn}$) von $D C_{np}$ reproduziert.
2. Ist C_{np} eine R_{nn} (bzw. $\bar{R}_{nn}$) reproduzierende Matrix, so wird Q_{nn} (bzw. $\bar{Q}_{nn}$) von $D^{-1} C_{np}$ reproduziert.

Beweis zu 1.:

C reproduziere Q ; es sei also eine positiv definite Matrix V vorgegeben, so daß gilt:

$$Q = C V C'$$

Multiplikation dieser Gleichung von links mit D und rechts mit D' ergibt wegen $D' = D$:

$$\begin{aligned} R &= D Q D = D(C V C')D' \\ &= (DC) V (DC)'. \end{aligned}$$

Also ist $(D C)$ eine Matrix, die entsprechend Definition (8) R reproduziert.

Beweis zu 2:

C reproduziere R ; sei also entsprechend Definition (8) eine positiv definite Matrix V vorgegeben, so daß gilt:

$$R = C V C'.$$

Da D^{-1} existiert und ebenfalls eine Diagonalmatrix ist, gilt $(D^{-1})' = D^{-1}$ und es folgt:

$$Q = D^{-1} R D^{-1} = D^{-1} C V C' D^{-1} = D(D^{-1} C) V (D^{-1} C)'.$$

Also ist $D^{-1} C$ eine Q reproduzierende Matrix.

Beweise für $\bar{Q}$ und $\bar{R}$: analog.

Dieser Satz bringt also zum Ausdruck, daß es - mathematisch gesehen - gleichgültig ist, ob man bei einer Faktorenanalyse von der Kovarianzmatrix oder der Korrelations-

matrix ausgeht. Die Ergebnisse sind äquivalent.

Kommunalitäten und reproduzierende Matrix.

In vielen Anwendungsfällen der Faktorenanalyse, etwa der Analyse von Korrelationskoeffizienten zwischen verschiedenen Tests, sind die Werte in der Hauptdiagonale der reduzierten Korrelationsmatrix, die sog. Kommunalitäten, nicht bekannt. Da diese jedoch bereits beim Beginn der Analyse benötigt werden, hilft man sich mit einer Schätzung der Kommunalitäten und verbessert diese Schätzungen iterativ. Zur Kommunalitätenschätzung sind verschiedene Verfahren in Gebrauch, die zu jeweils anderen Anfangs-Schätzwerten führen.

Abgesehen von der Forderung nach Konvergenz bei iterativer Verbesserung der Ausgangs-Schätzwerte ist ein solches Vorgehen nur dann statthaft, wenn die Kommunalitäten eindeutig bestimmt sind; denn von der Höhe der Diagonalelemente hängt ja das Aussehen einer Faktorenlösung schließlich auch ab.

Mit diesem Kommunalitätenproblem haben sich daher bereits die verschiedensten Autoren kritisch auseinandergesetzt. So wies schon LEDERMANN (1937) nach, daß bei gegebenen Korrelationskoeffizienten die Zahl der gemeinsamen Faktoren nicht eindeutig festzuliegen braucht. Zur vollständigen Reproduktion derselben Korrelationskoeffizienten kann man also u. U. unterschiedlich viele gemeinsame Faktoren heranziehen, so daß man auch zu unterschiedlichen Werten für die Kommunalitäten gelangt. Offenbar als Reaktion hierauf setzten andere Autoren (s. z. B. THURSTONE, 1947, S. 282-283) als Zahl der gemeinsamen Faktoren den kleinsten Rang fest, den die reduzierte Korrelationsmatrix der manifesten Variablen bei entsprechender Wahl der Hauptdiagonal-Elemente überhaupt annehmen kann.

HARMAN widmet dem Kommunalitätenproblem in seinem Buch "Modern factor analysis" (HARMAN, 1960) ein ganzes Kapitel,

in dem er u. a. Rechenverfahren angibt, mit deren Hilfe im Falle von höchstens zwei gemeinsamen Faktoren die Kommunalitäten der beobachteten Variablen - offenbar in eindeutiger Weise - ausgerechnet werden können. Die Lösung des Kommunalitätenproblems ist anscheinend ALBERT (1944) gelungen: er weist nach, daß - sofern die Zahl der beteiligten gemeinsamen Faktoren kleiner ist als die halbe Anzahl der eingehenden beobachteten Variablen - die Kommunalitäten algebraisch eindeutig bestimmt werden können.

Möglicherweise aber sehen sowohl ALBERT als auch HARMAN das Eindeutigkeitsproblem nicht ganz richtig: Es kommt nicht darauf an, ein mathematisches Verfahren zu entwickeln, das als Ergebnis genau einen Wert für jede Kommunalität liefert, sondern vielmehr darauf, daß bei vorgegebenen Interkorrelationen der beobachteten Variablen und vorgegebenem Rang der reduzierten Korrelationsmatrix - also vorgegebener Zahl der gemeinsamen Faktoren - jede Matrix, die die Interkorrelationen zu reproduzieren gestattet, auch zu denselben Kommunalitäten führt.

Die so verstandene Eindeutigkeit der Kommunalitäten ist aber tatsächlich nicht gegeben, wie durch das folgende Beispiel belegt wird:

Betrachtet seien die beiden Matrizen

$$A_{52} =: \begin{pmatrix} 0,8 & 0,6 \\ 0,3 & 0,5 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad C_{52} =: \begin{pmatrix} 0,3 & 0,6 \\ 0,8 & 0,5 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

und ein Satz von beobachtbaren Variablen $X_1, \dots, X_5$, deren Varianzen zu 1 normiert sind.

Diese Variablen mögen folgende Interkorrelationsmatrix besitzen:

$$\begin{pmatrix} & 0,54 & 0,6 & 0,6 & 0,6 \\ 0,54 & & 0,5 & 0,5 & 0,5 \\ 0,6 & 0,5 & & 1 & 1 \\ 0,6 & 0,5 & 1 & & 1 \\ 0,6 & 0,5 & 1 & 1 & \end{pmatrix}$$

Hierbei sind die Werte für die Kommunalitäten in der Haupt-Diagonale ausgespart.

Nimmt man nun an, daß den $X_1, \dots, X_5$ zwei unabhängige gemeinsame Faktoren G_1, G_2 mit Varianzen von 1 unterliegen und faßt zunächst A_{52} als Matrix von Ladungen auf diesen Faktoren auf, so muß $A_{52}A'_{25}$ obige Korrelationskoeffizienten richtig wiedergeben. Das ist auch in der Tat der Fall. Ersetzt man nun A_{52} durch C_{52} , so erhält man dieselben Korrelationskoeffizienten. Im ersten Fall erhält man aber als Kommunalitäten

$$1 \quad 0,34 \quad 1 \quad 1 \quad 1 ,$$

im zweiten Falle dagegen

$$0,44 \quad 0,89 \quad 1 \quad 1 \quad 1 .$$

Der Unterschied fällt ins Auge.

Zur Klärung der Frage, ob A_{52} und C_{52} linear ineinander transformiert werden können, nehme man an, es gäbe eine lineare Transformation L , die A_{52} in C_{52} überführt. Bei dieser Transformation müssen alle Punkte auf der Achse " G_2 " (s. unten stehende Abbildung 1) fest bleiben, also auch der Schnittpunkt S der durch die Punkte P_1 und P_2 gehenden Geraden mit der " G_2 "-Achse. Eine lineare Transformation bildet Punkte auf einer Geraden stets wieder in Punkten auf einer Geraden ab; im Widerspruch dazu liegen aber S und die zu P_1 und P_2 gehörenden Bildpunkte P'_1 und P'_2 nicht mehr auf einer Geraden.

A_{52} und C_{52} können also linear nicht ineinander transformiert werden.

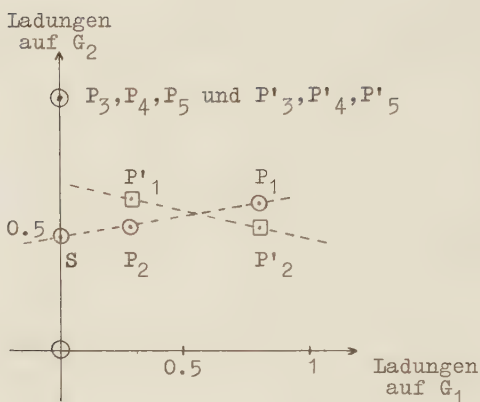


Abbildung 1: Graphische Darstellung der Ladungen auf den Faktoren G_1 und G_2 . P_1 bis P_5 beziehen sich auf die zeilenweise fortlaufend nummerierten Ladungspaare von A_{52} , P'_1 bis P'_5 auf die entsprechenden Ladungspaare von C_{52} .

Dieses Beispiel zeigt sehr deutlich, daß man zu unabhängigen Faktoren Ladungsmatrizen finden kann, mit deren Hilfe man zwar die Korrelationskoeffizienten einer Korrelationsmatrix richtig reproduzieren kann, die aber zu sehr unterschiedlichen Kommunalitäten führen und nicht einmal durch eine lineare Transformation - geschweige denn durch eine orthogonale - aufeinander bezogen werden können!

Aus diesem Grunde ist es unerlässlich, zu einer reduzierten Korrelationsmatrix bzw. Kovarianzmatrix stets diejenige Faktorenstruktur anzugeben, bezüglich derer die Reduktion der vollständigen Korrelationsmatrix erfolgt ist.

Bemerkung zu den unterschiedlichen Definitionen der reduzierten Kovarianzmatrix:

Auf Seite 42 wurde eine modifizierte Definition (6.1) der reduzierten Kovarianzmatrix diskutiert, bei der nicht nur die Varianzen, sondern auch die Kovarianzen zwischen den einzelnen spezifischen Faktoren berücksichtigt wurden. Bei durchgehender Benutzung dieser Definition kann man in den davon betroffenen Sätzen die Voraussetzung, daß die spezifischen Faktoren untereinander stochastisch unabhängig sein sollen, fallen lassen. Die Forderung der Unabhängigkeit der gemeinsamen von den spezifischen Faktoren reicht dann zum Beweis aus. Es sei bemerkt, daß sich dadurch der Nachweis zu einigen Behauptungen wesentlich vereinfacht.

Der faktorenanalytische Ansatz.

Ausgehend von der reduzierten Kovarianzmatrix wird bei einer Faktorenanalyse die gemeinsame Varianz der beobachteten Variablen nach Dimensionen aufgegliedert, sofern dies überhaupt möglich ist. Gefragt ist jedoch nach den Ladungen auf den gemeinsamen Faktoren des unterlegten Modells; damit diese aber zugänglich werden, müßte man von der doppelt reduzierten Kovarianzmatrix (s. Definition (6.2)) ausgehen.

Genau an dieser Stelle liegt der neuralgische Punkt der Faktorenanalyse. Sind nämlich die spezifischen Faktoren von den gemeinsamen unabhängig, kann man die gemeinsame Varianz für den Varianzanteil der gemeinsamen Faktoren nehmen; die reduzierte Kovarianzmatrix ist dann gleich der doppelt reduzierten und das Faktorenproblem wäre, wenigstens teilweise, gelöst. Sind aber die spezifischen Faktoren von den gemeinsamen abhängig, so liefern auch die spezifischen Faktoren einen Beitrag sowohl zur gemeinsamen Varianz der beobachteten Variablen als auch zu den Kovarianzen zwischen ihnen, so daß reduzierte und doppelt reduzierte Kovarianzmatrix nicht mehr übereinstimmen.

Empirisch zugänglich sind nun die Kovarianzen bzw. Korrelationskoeffizienten zwischen den manifesten Variablen und allenfalls noch deren Kommunalitäten. Doch schon diese Kommunalitäten müssen mit Hilfe mehr oder weniger komplizierter Verfahren geschätzt werden. Daß hierbei tatsächlich Kommunalitäten und nicht etwa Kommensurabilitäten (s. Definition (4.3)) geschätzt werden, erkennt man daran, daß die einschlägigen Schätzmethoden sämtlich von den vorliegenden Kovarianzen bzw. Korrelationskoeffizienten zwischen den Variablen ausgehen.

Um zur doppelt reduzierten Kovarianzmatrix zu gelangen, müßte man also vor Analysenbeginn neben den Kommunalitäten auch noch den Grad der Abhängigkeit zwischen spezifischen und gemeinsamen Faktoren kennen. Da Faktoren aber der direkten Beobachtung nicht zugänglich sind, können quantitative Angaben über sie nur über Faktorenanalysen gewonnen werden; eigens zu diesem Zwecke wurde diese Methode ja entwickelt. Insbesondere gilt das auch für spezifische Faktoren!

Vom Konzept der Faktorenanalyse her besteht nun kein Grund, unabhängigen gemeinsamen Faktoren den Vorzug vor abhängigen zu geben. Dem trägt man bekanntlich durch Zulassung und Anwendung von schiefwinkligen Rotationen Rechnung. Da nun die Einteilung in spezifische und gemeinsame Faktoren vom Rahmen der gerade in Betracht stehenden manifesten Variablen abhängt (s. dazu auch die Ausführungen auf S. 23), muß man akzeptieren, daß spezifische Faktoren untereinander und von gemeinsamen Faktoren abhängig sein können.

Man kann nun versucht sein, dieses Problem per definitionem zu lösen, indem man die Unabhängigkeit der Faktoren einfach unterstellt. Wegen der Unbeobachtbarkeit der Faktoren läßt sich diese Unterstellung nicht widerlegen. In den noch folgenden Ausführungen jedoch wird u. a. gezeigt, daß dieser - an sich schon nicht unbedenkliche Ausweg - auch prinzipiell versperrt ist: Infolge von Selektionen

und/oder Vorgängen, die Selektionen analog sind und sich mit denselben Hilfsmitteln beschreiben lassen, ist damit zu rechnen, daß Faktoren in der Regel abhängig sind, und zwar auch dann, wenn in irgendwelchen fiktiven Ausgangspopulationen deren Unabhängigkeit gegeben sein sollte.

Die Faktorenanalyse ist also im allgemeinen nicht in der Lage, die Ladungen auf den gemeinsamen Faktoren - also die Parameter des im Ansatz unterlegten Modells - zu bestimmen.

C. Die Selektion.

1. Selektion und Wahrscheinlichkeitsdichte.

Aus einer Stichprobe A, die man einer Population P von Individuen (besser: Meßwertträgern) entnommen hat, werde nach bestimmten Gesichtspunkten eine zweite Stichprobe B zusammengestellt. Wenn diese Gesichtspunkte ganz oder auch nur teilweise von den bei A erhaltenen Meßwerten abhängen, handelt es sich um eine "Selektion".

Die Stichprobe B ist ebenfalls eine Stichprobe aus der Population P; allerdings ist sie, sofern Selektion vorliegt, nicht repräsentativ; denn die Wahrscheinlichkeit, in die Stichprobe B zu gelangen, ist nicht für alle Individuen der Population P gleich, sondern hängt eben ab von den Meßwerten, die ein Individuum hinsichtlich der in

Betracht stehenden Größen besitzt.

Selektion bedeutet also die Veränderung der (mehrdimensionalen) Wahrscheinlichkeitsdichte der betroffenen zufälligen Variablen.

Indem sich nun die Dichte der von der Selektion unmittelbar betroffenen zufälligen Variablen ändert, wird die Dichte anderer, zunächst nicht "beachteter" Variablen, sofern sie von den ersteren stochastisch abhängen, mittelbar ebenfalls verändert.

Die Wahrscheinlichkeitsdichten gehen aber in die Formeln zur Berechnung der Varianzen von zufälligen Variablen und der Kovarianzen zwischen ihnen ein. Man wird daher erwarten, daß ein Selektionsvorgang Varianzen und Kovarianzen ebenfalls verändert.

Im folgenden sollen diese Überlegungen präzisiert werden.

Grundlegende Definitionen.

Gegeben sei eine n -dimensionale zufällige Variable U mit der Verteilungsdichte $f(u)$.

M sei die Menge der Dichten $f(x)$ aller möglichen Teilvariablen X von U ($f(u)$ eingeschlossen). Diese Dichten erhält man als Rand-Verteilungsdichten aus $f(u)$.

M^* sei eine Menge von zunächst nicht näher spezifizierten Funktionen (von 1 bis n Argumenten), die als Verteilungsdichten aufgefaßt werden können.

Sei a eine Abbildung von M in M^* ; für eine gegebene Teilvariable X von U werde mit $f^*(x)$ das durch a vermittelte Bild von $f(x)$ in M^* bezeichnet.

(14) Definition der Selektion:

a heiße "Selektion" bezüglich U , wenn für alle Teilvariablen X von U gilt: $f^*(x)$ erhält man als Randverteilungsdichte aus $f(u)$.

Damit ist für alle Elemente von M das zugehörige a -Bild im M^* bestimmt.

Selektion ist also definiert als eine Operation, die auf die Verteilungsdichten von zufälligen Variablen, nicht auf die Variablen selbst, angewendet wird; diese Verteilungsdichten (also die Elemente von M) werden durch Selektion in bestimmter Weise verändert.

(15) Definition der Selektionsvariablen:

X, Y seien zufällige Variablen; sei $U =: (X, Y)$. Sei a Selektion bezüglich U ; mit $f^*(x)$ werde das a -Bild von $f(x)$ bezeichnet.

X heiße "Selektionsvariable" bezüglich (a, U) , wenn gilt:

$$f^*(u) = f^*(x) \cdot f_x(y)$$

($f_x(y)$ ist bedingte Dichte, s. (1)).

Ersichtlich ist U Selektionsvariable auch bezüglich (a, U) , da in diesem Falle $f_u(u) = 1$ gilt.

Diese Definition trägt der Tatsache Rechnung, daß bei der Selektion die zufällige Variable $Y = (Y_1, \dots, Y_p)$ nur "mittelbar" insofern betroffen ist, daß nur bestimmte Realisationen x von X "bevorzugt" werden, darüber hinaus die freie Variierbarkeit von Y aber nicht eingeschränkt wird. D. h. es wird angenommen, daß die Dichte $f_x(y)$ von Y unter der Hypothese, daß bereits eine bestimmte Realisation x von X vorliegt, sich unter der Selektion nicht ändert.

Für den Fall also, daß die Selektionsvariable X von geringerer Dimension als U ist, braucht man nur zu wissen, daß $f(x)$ durch die Selektion in $f^*(x)$ übergeführt wird, um die Auswirkungen der Selektion auf alle übrigen interessierenden Wahrscheinlichkeitsdichten - im Prinzip wenigstens - berechnen zu können.

Bemerkung:

Bei den hier gegebenen Definitionen der Selektion (14) und der Selektionsvariablen (15) handelt es sich um die mathematische Beschreibung von tatsächlichen Auslesevorgängen. Diese Definitionen stimmen mit von anderen Autoren (als Beispiele seien lediglich LAWLEY, 1943, S. 28; BIRNBAUM et al., 1950, S. 193 genannt) gegebenen praktisch überein, obwohl ein prinzipieller Unterschied besteht: Die Definition der Selektionsvariable wird anderswo offenbar als Definition der Selektion selbst angesehen, was unzumutbar ist. Auf diese Weise nämlich wird der Blick für einige Eigenschaften von Selektionsvariablen, die sehr einfach zu formulieren sind, verstellt. Gerade diese Eigenschaften, ausgedrückt in den folgenden drei Selektionssätzen, erweisen sich als für die Theorie der Selektion hinsichtlich der Faktorenanalyse von entscheidender Bedeutung.

Einfache Selektionssätze.

(16) Satz:

X, Y, Z seien zufällige Variable.

- (a) Sei a eine Selektion bezüglich (X, Y, Z) und X Selektionsvariable bezüglich $(a, (X, Y, Z))$.

Behauptung:

X ist Selektionsvariable auch bezüglich $(a, (X, Y))$.

Beweis:

Es gilt:

$$\begin{aligned} f^*(x, y, z) &= f^*(x) \cdot f_x(y, z) \\ (15) \\ &= f^*(x) \cdot f_x(y) \cdot f_{(x, y)}(z) \\ (1.3) \end{aligned}$$

Dann wird

$$\begin{aligned}
 f^*(x,y) &= \int_{-\infty}^{+\infty} f^*(x,y,z) \, dz = \int_{-\infty}^{+\infty} f^*(x) \cdot f_x(y) \cdot f_{(x,y)}(z) \, dz \\
 &= f^*(x) \cdot f_x(y) \cdot \underbrace{\int_{-\infty}^{+\infty} f_{(x,y)}(z) \, dz}_{= 1 \text{ (da } f_{(x,y)}(z) \text{ Dichte ist)}} \\
 &= f^*(x) \cdot f_x(y)
 \end{aligned}$$

Entsprechend Definition (15) ist damit X Selektionsvariable bezüglich (a, (X,Y)).

Daß eine Selektionsvariable jedoch nicht eindeutig bestimmt zu sein braucht, besagt folgender

(17) Satz:

Unter denselben Gegebenheiten wie in (16) gilt die

Behauptung:

(X,Y) ist ebenfalls Selektionsvariable bezüglich (a, (X,Y,Z)).

Beweis:

Nach (16) und (15) ist

(b) $f^*(x,y) = f^*(x) \cdot f_x(y)$. Ferner gilt:

$$\begin{aligned}
 f^*(x,y,z) &= f^*(x) \cdot f_x(y,z) \\
 (15)(a) \\
 &= f^*(x) \cdot f_x(y) \cdot f_{(x,y)}(z) \\
 (1.3) \\
 &= f^*(x,y) \cdot f_{(x,y)}(z) \\
 (b)
 \end{aligned}$$

Hierauf Definition (15) angewendet ergibt die Behauptung.

Damit wird ersichtlich, daß alle Variablen, die X umfassen, ebenfalls Selektionsvariable im Sinne der Definition sind; bei Abgrenzung einer speziellen Selektionsvariablen kann also u. U. recht willkürlich vorgegangen werden.

Es erscheint daher als absurd, bei statistischen Analysen, wie etwa der Faktorenanalyse, die Selektionsvariablen von der Bearbeitung ausschließen zu wollen.

Der Frage, ob es unter den Selektionsvariablen eine kleinsten Dimension gibt, soll hier nicht weiter nachgegangen werden. Es sei jedoch darauf hingewiesen, daß man oft schon unter ganz gewöhnlichen Umständen (s. die folgenden Beispiele) eine Variable ausfindig machen kann, die entsprechend der Definition (15) die betrachtete Selektion vollständig bestimmt, die aber nicht zu den ursprünglich betrachteten Ausgangsvariablen gehört.

Verteilungsdichten von Variablen, die von einer Selektionsvariablen unabhängig sind, werden von der Selektion auch nicht betroffen:

(18) Satz:

X, Y seien zufällige Variable.

Sei a Selektion bezüglich (X, Y) und sei X Selektionsvariable bezüglich $(a, (X, Y))$.

Voraussetzung:

$$f(x, y) = f(x) \cdot f(y) \quad (\text{d. h.: } X, Y \text{ sind stochastisch unabhängig})$$

Behauptungen:

$$(18.1) \quad f^*(y) = f(y)$$

$$(18.2) \quad f^*(x, y) = f^*(x) \cdot f(y), \text{ d. h. die Unabhängigkeit von } X, Y \text{ bleibt auch nach der Selektion erhalten.}$$

Beweis zu (18.1):

Es gilt:

$$f^*(x,y) = f^*(x) \cdot f_x(y) \quad (\text{Definition (15)}).$$

Wegen der vorausgesetzten Unabhängigkeit von X,Y gilt aber

$$f_x(y) = f(y), \quad \text{also folgt}$$

$$(a) \quad f^*(x,y) = f^*(x) \cdot f(y). \quad \text{Somit wird:}$$

$$\begin{aligned} f^*(y) &= \int_{-\infty}^{+\infty} f^*(x,y) dx = \int_{-\infty}^{+\infty} f^*(x) \cdot f(y) dx = f(y) \cdot \underbrace{\int_{-\infty}^{+\infty} f^*(x) dx}_{=1 \text{ (da Dichte)}} \\ &= f(y). \end{aligned}$$

Der Beweis zu (18.2) ist in der mit (a) bezeichneten Zeile des Beweises zu (18.1) zu finden.

Univariate und multivariate Selektion.

In der Literatur wird eine Selektion als univariat bezeichnet, wenn die entsprechende Selektionsvariable eindimensional ist. Analog versteht man unter multivariater Selektion eine solche, bei der die Selektionsvariable mehrdimensional ist.

Diese Einteilung ist jedoch problematisch, da die Dimensionalität der Selektion nach Satz (17) nicht eindeutig feststeht und man häufig, wie sich in den gleich folgenden Beispielen zeigen wird, durch Erweiterung des Systems der betrachteten zufälligen Variablen aus einer multivariaten eine univariate Selektion machen kann. Umgekehrt kann durch Streichung einer Variablen eine bisher univariate Selektion multivariat werden.

Dieser Unterscheidung dürfte demnach keine besondere theoretische Bedeutung zukommen.

2. Beispiele für spezielle Arten von Selektionen.

Betrachtet seien zwei zufällige Variable X_1, X_2 , abgekürzt geschrieben als X . Mit $f(x_1, x_2)$ werde die Wahrscheinlichkeitsdichte von $X = (X_1, X_2)$ vor Selektion und mit $f^*(x_1, x_2)$ die Wahrscheinlichkeitsdichte nachher bezeichnet. u, v seien zwei fest vorgegebene Zahlen mit $0 < u < v$.

Beispiel 1:

Die Selektion a sei wie folgt vorgegeben:

$$f^*(x_1, x_2) =: \begin{cases} c f(x_1, x_2) & \text{für } u \leq x_1 + x_2 \leq v \\ 0 & \text{sonst.} \end{cases}$$

c dient zur Normierung und wird so bestimmt, daß

$$\int_{-\infty}^{+\infty} f^*(x) dx = 1 \text{ wird.}$$

Bei Veranschaulichung dieser Selektion in einem zweidimensionalen Koordinatensystem sieht man, daß das Gebiet der Ebene, in der $f \neq 0$ sein kann, ein gegen die x_1 -Achse um -45° geneigter Streifen der Breite $b = (v-u)/\sqrt{2}$ ist. (s. Abbildung 2).

In diesem Beispiel ist $X = (X_1, X_2)$ Selektionsvariable bzgl. $(a, (X_1, X_2))$. Demnach müßte es sich bei a um eine bivariate Selektion handeln. Man kann die Selektion a aber auch wie folgt vorgeben:

Sei $X_0 =: X_1 + X_2$ und $f(x_0)$ die Dichte von X_0 . Dann ist durch

$$f^*(x_0) =: \begin{cases} c f(x_0) & \text{für } u \leq x_0 \leq v \\ 0 & \text{sonst} \end{cases}$$

dieselbe Selektion wie oben definiert. Ersichtlich ist aber nun die eindimensionale Variable X_0 Selektionsvariable bzgl. $(a, (X_0, X_1, X_2))$, so daß man a auch als univariate Selektion auffassen kann.

Die Eigenschaft, Selektionsvariable zu sein, kommt also einer Variablen nicht "automatisch" durch den speziellen Selektionsvorgang zu, der durch sie beschrieben wird, sondern ist vielmehr abhängig von dem Bezugsrahmen der übrigen - sozusagen "konkret hingeschriebenen" - Variablen. Dasselbe gilt für die Eigenschaft einer Selektion, univariat oder multivariat zu sein.

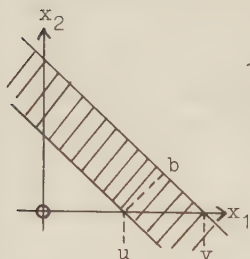


Abbildung 2.

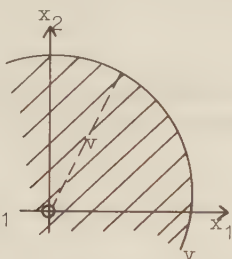


Abbildung 3.

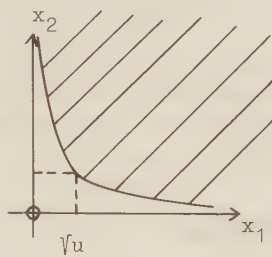


Abbildung 4.

Graphische Darstellungen der in den Beispielen 1, 2 und 3 beschriebenen Selektionen. Der Bereich, in dem die Wahrscheinlichkeitsdichte nach Selektion ungleich Null sein kann, in dem also Realisationen von (X_1, X_2) möglich sind, ist schraffiert.

Beispiel 2:

Eine Selektion a sei gegeben durch:

$$f^*(x_1, x_2) =: \begin{cases} c f(x_1, x_2) & \text{für } x_1^2 + x_2^2 \leq v^2 \\ 0 & \text{sonst.} \end{cases}$$

Die möglichen Realisationen von $X = (X_1, X_2)$ liegen innerhalb einer Kreisfläche mit dem Mittelpunkt O und dem Radius v . (s. Abbildung 3).

Durch die Vorschriften

$$X_0 =: X_1^2 + X_2^2 \quad \text{und}$$

$$f^*(x_0) =: \begin{cases} c \cdot f(x_0) & \text{für } x_0 \leq v^2 \\ 0 & \text{sonst} \end{cases}$$

ist a ebenfalls bestimmt. Auch hier kann man, je nach dem verwendeten Bezugsrahmen, a als bivariate oder univariate Selektion auffassen.

In der Praxis kommen solche Selektionen vor, wenn eine Gruppe von Individuen so zusammengestellt wird, daß bezüglich aller in Betracht stehenden Merkmale relative Homogenität vorliegt.

Beispiel 3:

Unter der Voraussetzung $X_1 \geq 0$ sei eine Selektion a gegeben durch:

$$f^*(x_1, x_2) =: \begin{cases} c \cdot f(x_1, x_2) & \text{für } u \leq x_1 \cdot x_2 \\ 0 & \text{sonst.} \end{cases}$$

Mögliche Realisationen von (X_1, X_2) liegen innerhalb eines Hyperbelastes, der durch den Punkt $(\sqrt{u}, \sqrt{u})$ geht und der die Koordinatenachsen nicht schneidet (s. Abbildung 4).

Auch durch $X_0 =: X_1 \cdot X_2$ und

$$f^*(x_0) =: \begin{cases} c \cdot f(x_0) & \text{für } u \leq x_0 \\ 0 & \text{sonst} \end{cases}$$

ist a definierbar.

Mit solchen Selektionen muß man rechnen bei Variablen etwa aus dem Leistungsbereich, wenn der Ansatz

$$\text{Leistung} \triangleq \text{Fähigkeit mal Motivation}$$

Gültigkeit hat und die Leistungsbesten ausgesucht werden.

Beispiel 4:

In den bisherigen Beispielen waren die betrachteten Selektionsvorgänge Trunkationen. Setzt man jedoch die Auftretenswahrscheinlichkeit von denjenigen Realisationen, die bestimmten Bedingungen nicht genügen, nicht gleich Null, sondern modifiziert sie lediglich, so liegt probalistische Selektion vor. Probalistisch kann auf mancherlei Art selegiert werden; hier seien nur zwei einfache Arten aufgezeigt:

1. X_0 sei eine zufällige Variable, die eine Funktion von X_1 und X_2 ist: $X_0 =: F(X_1, X_2)$. Dann ist folgende Selektion denkbar:

$$f^*(x_0) =: \begin{cases} c f(x_0) & \text{für } u \leq x_0 \leq v \\ c/2 f(x_0) & \text{sonst.} \end{cases}$$

2. Z_1 sei eine weitere eindimensionale Variable mit dem Erwartungswert Null. Dann ist durch

$$Y =: X_0 + Z_1$$

$$f^*(y) =: \begin{cases} c f(y) & \text{für } u \leq y \leq v \\ 0 & \text{sonst.} \end{cases}$$

ebenfalls eine hinsichtlich X_1, X_2 probalistische Selektion gegeben.

Dieselbe Selektion kann auch auf folgende Weise definiert werden:

$$f^*(x_1, x_2) =: \begin{cases} c f(x_1, x_2) & \text{für } u \leq x_0 + z \leq v \\ 0 & \text{sonst.} \end{cases}$$

Beispiel 5:

Ein etwas andersartiger Selektionstyp kann vereinfacht wie folgt beschrieben werden:

X sei eine zufällige Variable mit beliebiger Verteilungsform.

Sei

$$f^*(x) =: \begin{cases} c = \text{const.} & \text{für } u \leq x \leq v \\ 0 & \text{sonst.} \end{cases}$$

Hier erfolgt also die Selektion so, daß innerhalb eines gewissen Bereiches alle möglichen Werte von X mit gleicher Wahrscheinlichkeit auftreten. Durch solche Selektionen werden also die normalerweise meist geringen Häufigkeiten der Verteilungsenden "künstlich" angehoben, während die mittleren Bereiche "ausgedünnt" werden.

Solche Selektionen können bei der Testkonstruktion vorkommen, wenn man bestrebt ist, eine gleich gute Differenzierungsfähigkeit des Tests über einen großen Eigenschaftsbereich zu erzielen. In diesem Falle nämlich muß man die Testitems auf die verschiedenen Schwierigkeitsbereiche gleichmäßig verteilen.

Ferner können solche Selektionen auch bei Skalierungsexperimenten vorkommen, wenn die Stimuli hinsichtlich ihres Ausprägungsgrades nicht mehr mit derjenigen Häufigkeit dargeboten werden, mit der sie unter normalen Umständen auftreten, sondern jede Abstufung gleich häufig unter Ausschluß von extremeren Werten geboten wird.

Beispiel 6:

In den bisherigen Beispielen zielten die Selektionen im wesentlichen darauf ab, die Variabilität bzw. Varianz der betroffenen Variablen zu verändern. Es sind aber auch solche Selektionen denkbar, die primär die Kovarianzen zwischen Variablen, weniger deren Varianzen, beeinflussen. Zur Veranschaulichung dieses Sachverhaltes diene das folgende Beispiel, das aber darüber hinaus noch verdeutlichen soll, daß Selektionen als Veränderung von Auftretenswahrscheinlichkeiten definiert sind. Solche Veränderungen können, müssen aber nicht durch Auslesevorgänge im wörtlichen Sinne hervorgerufen werden.

Man denke sich einen Beurteiler vor die Aufgabe gestellt, eine Menge von N Bekannten hinsichtlich mehrerer Eigenschaftsbezeichnungen, z. B. Hilfsbereit, Verlässlich, Fleißig sowie Rechthaberisch, Ehrgeizig und Beherrscht zu beurteilen. Die Abgabe der Beurteilungen soll durch Ankreuzen auf einer etwa 7-stufigen Skala, die jeder Eigenschaftsbezeichnung beigegeben ist, erfolgen, und zwar getrennt für jeden einzelnen Bekannten. *)

Bei solchen Beurteilungen ist mit dem Auftreten des sog. "Halo"-Effektes, d. h. mit der Begünstigung gleichartiger Beurteilungen hinsichtlich verschiedener Eigenschaften, zu rechnen: Beurteilt der Beurteiler z. B. gerade einen ihm sympathischen Menschen, so ist er geneigt, ihm - ohne Rücksicht auf dessen tatsächliches Verhalten - positiv bewertete Eigenschaften generell zuzuschreiben und negativ bewertete generell abzusprechen. Hinsichtlich eines unsympathischen Bekannten geschieht das Gegenteil.

Die Variablen nun, die hier in Betracht stehen, sind die abzugebenden Beurteilungen hinsichtlich der sechs Eigenschaftsbezeichnungen. Die jeweils sechs Beurteilungen, die der Beurteiler bezüglich eines speziellen Bekannten abgibt, sind als Realisationen dieser Variablen aufzufassen. Durch spezielle Wahl der Bekannten - indem man den Beurteiler z. B. nur sympathische und unsympathische oder nur gleichgültige Bekannte beurteilen läßt - kann man also die Kovarianzen zwischen den Eigenschaftsbezeichnungen innerhalb gewisser Grenzen manipulieren.

Beispiel 7:

Schließlich sei noch ein weiteres Beispiel dafür gebracht, daß sich ein spezieller Selektionsvorgang vorwiegend auf die Kovarianzen zwischen Variablen auswirken kann.

*) Die genannten Eigenschaftsbezeichnungen wurden in Untersuchungen von MERZ (1957) und in einer Arbeit des Verfassers selbst (KALVERAM, 1962) in ähnlicher Weise wie oben beschrieben verwendet.

Man stelle sich vor, es solle ein Persönlichkeitstest, bestehend aus etwa 100 Einzelfragen, konstruiert werden. Dazu stellt man zunächst eine Liste von rund 200 Items zusammen, die man einer Gruppe von Versuchspersonen zur Beantwortung vorlegt. In der sich anschließenden Aufgabenanalyse wird man solche Items als für den Test ungeeignet zurückweisen, die von allen Versuchspersonen in etwa gleicher Weise beantwortet wurden (z. B. von fast allen mit "Ja"). Diese Items nämlich können mangels Varianz nicht zwischen Probanden differenzieren.

In weitergehenden statistischen Analysen berechnet man häufig nicht die Korrelationen zwischen Items, sondern die Korrelationen zwischen Versuchspersonen. Hierbei werden die Antworten einer Versuchsperson aufgefaßt als Realisation der ihr zugeordneten Verhaltensvariable (sog. Q-Technik, verwendet zur Bestimmung von "Typen"). Die Entfernung der vorwiegend gleich beantworteten Items aus dem Test hat eine Verminderung der Korrelationen zwischen den Versuchspersonen notwendig zur Folge.

Items mit großer Varianz der abgegebenen Antworten wird man hingegen bei der Konstruktion des Tests bevorzugt beibehalten. Das aber muß - wenigstens im Durchschnitt - zu negativen Kovarianzen zwischen den Versuchspersonen führen, wie man an Hand der folgenden, stark vereinfachten Hilfsüberlegung einsehen kann: Im Falle von nur zwei Versuchspersonen und maximaler Varianz aller Items des Tests muß die eine Versuchsperson mit "Ja" geantwortet haben, wenn die andere "Nein" als Antwort gegeben hat. D. h. beide Versuchspersonen beantworten in diesem Falle alle Items gegenläufig und korrelieren somit zu -1 miteinander.

Bezogen auf die sogenannte invertierte Korrelations-technik also sind durch die Itemselektion, die bei der Konstruktion des Tests vorgenommen worden ist, die Kovarianzen zwischen den Versuchspersonen in die negative Richtung verschoben, wohingegen die Varianz innerhalb der Antworten einer Versuchsperson kaum betroffen sein dürfte.

Möglicherweise trägt der Umstand, daß verschiedene Persönlichkeitstests in ähnlicher oder gleicher Weise nach objektiven Kriterien entwickelt wurden, mit dazu bei, daß man bei Faktorenanalysen (Q-Technik) solcher Tests nicht selten zu "ähnlichen" Ergebnissen kommt.

3. Selektion und Kovarianz.

Generell zu erwartende Veränderungen.

$X = (X_1, X_2, \dots, X_n)$ sei eine zufällige Variable. Mit $X' \circ X$ werde die Kovarianz-Matrix von X bezeichnet.

Unter der Wirkung einer Selektion bezüglich X geht $f(x)$ in $f^*(x)$ über. Damit ändern sich auch die Dichten aller ein- und zweidimensionalen Teilvariablen von X , d. h. i. a. gilt:

$$f^*(x_i) \neq f(x_i) \quad (i = 1, 2, \dots, n) \quad \text{und}$$

$$f^*(x_i, x_j) \neq f(x_i, x_j) \quad (i, j = 1, 2, \dots, n; i \neq j).$$

Über diese Dichten sind aber die Varianzen der zufälligen Variablen $X_1, \dots, X_n$ und die Kovarianzen zwischen ihnen definiert. Somit wird die Kovarianz-Matrix $X' \circ X$ von X infolge der Selektion übergehen in eine davon verschiedene Matrix $(X' \circ X)^*$.

Bei Kenntnis von $f^*(x)$ kann man zu Aussagen über $(X' \circ X)^*$ gelangen; es zeigt sich jedoch, daß man zu Aussagen über diese Matrix gelangen kann auch ohne detaillierte Kenntnis von $f^*(x)$:

Die Selektionsformeln von LAWLEY.

Unter den Voraussetzungen, daß erstens die Kovarianzmatrix der Selektionsvariablen positiv definit ist und zweitens die Regression der von einer Selektion nicht

direkt betroffenen Variablen auf die Selektionsvariablen linear und homoskedastisch ist, kann man nach LAWLEY (1943) den Selektionseffekt auf alle beteiligten Variablen schon dann berechnen, wenn man nur die Auswirkungen der Selektion auf die Selektionsvariablen allein kennt. Hierbei ist die spezielle Art der Selektion unerheblich.

Die LAWLEYSchen Selektionsformeln lauten wie folgt: Betrachtet sei eine n-dimensionale zufällige Variable $X = (X_1, \dots, X_p, X_{p+1}, \dots, X_{p+q})$ ($n = p + q$). ($X_1, \dots, X_p$) sei Selektionsvariable bezüglich X .

Die Kovarianzmatrix V_{nn} von X sei wie folgt aufgespalten:

$$V_{nn} = \left(\begin{array}{c|c} V_{pp} & V_{pq} \\ \hline V_{qp} & V_{qq} \end{array} \right) .$$

Hierbei bedeuten V_{pp} die Kovarianzmatrix der p Selektionsvariablen $X_1, \dots, X_p$ und V_{qq} die Kovarianzmatrix der q übrigen, nur indirekt von der Selektion betroffenen Variablen $X_{p+1}, \dots, X_{p+q}$.

Durch die Selektion geht die so unterteilte Kovarianzmatrix V_{nn} über in

$$V_{nn}^* = \left(\begin{array}{c|c} V_{pp}^* & V_{pq}^* \\ \hline V_{qp}^* & V_{qq}^* \end{array} \right) .$$

Bei Kenntnis des Aussehens von V_{pp}^* kann man nun die Teilmatrizen V_{pq}^* , V_{qp}^* , V_{qq}^* von V_{nn}^* berechnen:

$$(19) \quad \begin{aligned} V_{pq}^* &= V_{pp}^* V_{pp}^{-1} V_{pq} ; & V_{qp}^* &= V_{qp} \\ V_{qq}^* &= V_{qq} - V_{qp} (V_{pp}^{-1} - V_{pp}^{-1} V_{pp}^* V_{pp}^{-1}) V_{pq}' \end{aligned}$$

Kompensatorische Kovarianz als spezieller Selektionseffekt.

Mit Hilfe der LAWLEYSchen Formeln läßt sich nun schnell zeigen, welche Auswirkungen eine Selektion auf zufällige Variable haben muß, wenn die Varianz der Summenvariable eingeschränkt wird:

Gegeben seien die zufälligen Variablen $X_1, \dots, X_n$.

(a) Sei $X_0 =: a_1 X_1 + a_2 X_2 + \dots + a_n X_n$.

X_0 ist also eine gewichtete Summe der $X_1, \dots, X_n$.

X_0 sei nun aufgefaßt als Selektionsvariable, deren Varianz σ^2 durch eine Selektion (z. B. Trunkation) auf den Wert σ^{*2} ($\sigma^{*2} < \sigma^2$) eingeschränkt werden soll.

Die Kovarianzmatrix von $(X_0, X_1, \dots, X_n)$ hat folgendes Aussehen:

$$V_{nn} = \begin{pmatrix} \sigma^2 & a_1 & a_2 & \dots & a_n \\ a_1 & & & & \\ a_2 & & & & \\ \vdots & & & & \\ a_n & & & & \end{pmatrix} \begin{matrix} \\ \\ \\ \\ V_{qq} \end{matrix}.$$

Es ist also hier $p = 1$, $q = n$ und $V_{pq} = (a_1, \dots, a_n)$. V_{pp} besteht nur aus der Zahl der σ^{*2} .

In diesem Falle erhält man für die Kovarianzmatrix V_{qq}^* der Variablen $X_1, \dots, X_n$ nach LAWLEY (s. (19))

$$9.1) \quad V_{qq}^* = V_{qq} - \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} (k) (a_1, \dots, a_n),$$

wobei $k =: \frac{1}{\sigma^2} - \frac{1}{\sigma^2} \sigma^{*2} \cdot \frac{1}{\sigma^2} = \frac{1}{\sigma^2} (1 - \frac{\sigma^{*2}}{\sigma^2}) > 0$ ist.

Ausgeschrieben für die Variablen X_i und X_j erhält man also

$$9.2) \quad (X_i \circ X_j)^* = (X_i \circ X_j) - a_i a_j \cdot k.$$

D. h. also, die Kovarianz zwischen X_i und X_j nach Selektion erhält man, indem man von der ursprünglichen Kovarianz zwischen den beiden Variablen einen Term abzieht. Es ist nichts im Wege, diesen Term die - infolge des Selektionsvorganges aufgetretene - kompensatorische Kovarianz zu nennen.

"Ausgeartete" Selektionen.

Betrachtet sei eine zufällige Variable $X =: (X_1, X_2, \dots, X_n)$. Es ist nun denkbar, daß zwischen den Komponenten von X eine lineare Beziehung der folgenden Art besteht:

$$(20) \quad c_1 X_1 + c_2 X_2 + \dots + c_n X_n = k = \text{const.}$$

($c_1, \dots, c_n$ reell und nicht zusammen gleich Null.)

Durch Umformen erhält man z. B.

$$X_1 = -1/c_1 (c_2 X_2 + \dots + c_n X_n - k),$$

d. h. X_1 ist vollständig determiniert durch die übrigen $n-1$ zufälligen Variablen.

Einfachstes Beispiel für das Bestehen einer solchen linearen Beziehung ist $X_2 = X_1$.

Indem man nun bei Durchführung einer Selektion für gegebene Zahlen $c_1, \dots, c_n, k$ nur solche Realisationen der $X_1, \dots, X_n$ zuläßt, die die Beziehung (20) erfüllen, kann man eine lineare Beziehung zwischen den $X_1, \dots, X_n$ "künstlich" erzeugen, auch wenn sie vorher nicht bestanden hat.

Es ist nun zweckmäßig, schon dann von einer linearen Beziehung zwischen zufälligen Variablen zu sprechen, wenn folgendes gilt:

(21) Es gibt Zahlen $c_1, \dots, c_n$, die nicht zusammen gleich Null sind, so daß die "neue" Variable

$$Y =: c_1 X_1 + \dots + c_n X_n$$

die Varianz Null besitzt.

Daher sei definiert:

(22) Definition:

Eine Selektion a bezüglich X heiße ausgeartet, wenn durch sie die Zahl der vor der Selektion bestehenden linearen Beziehungen der Art (21) vergrößert wird, sie heiße nicht-ausgeartet, wenn diese Zahl sich nicht erhöht.

(23.1) Es gilt nun:

Wenn es zwischen den $X_1, \dots, X_n$ keine linearen Beziehungen der Art (21) gibt, so ist die Kovarianzmatrix Q von X positiv definit.

(Zum Beweis dieser Behauptung siehe etwa bei SCHMETTERER, S. 111).

Umgekehrt gilt auch:

(23.2) Wenn die Kovarianzmatrix Q von X positiv definit ist, gibt es zwischen den Komponenten von X keine linearen Beziehungen der Art (21).

Beweis zu (23.2):

(a) Sei Q als positiv definit vorgegeben.

Angenommen nun, es gäbe eine lineare Beziehung der Art (21); dann seien Zahlen $c_1, c_2, \dots, c_n$ vorgegeben, die nicht zusammen gleich Null sind und für die

(b) $Y_1 =: c_1 X_1 + \dots + c_n X_n$

die Varianz Null besitzt.

Eine Matrix C_{nn} sei wie folgt definiert:

$$C_{nn} =: \begin{pmatrix} c_1 & c_2 & \dots & c_n \\ x & x & \dots & x \\ \vdots & \vdots & \dots & \vdots \\ x & x & \dots & x \end{pmatrix},$$

wobei an den mit x bezeichneten Stellen die Werte so gewählt seien, daß C_{nn} regulär ist.

Eine neue zufällige Variable sei nun definiert durch

(c) $Y' =: CX'$

Die erste Komponente von Y ist Y_1 , definiert unter (b). Die Kovarianzmatrix $\bar{Q} =: Y' \circ Y$ von Y besitzt daher eine Hauptdiagonale, in der mindestens ein Element verschwindet. Somit kann nach (2.6) $\bar{Q}$ nicht positiv definit sein.

ren besteht. Es ist bereits darauf hingewiesen worden (s. S.33), daß diese Grundgleichungen lediglich den bei allen Individuen einer definierten Population unterstellten gleichartigen Zusammenhang der in Betracht stehenden Größen abgekürzt zum Ausdruck bringen.

Dieser Zusammenhang gilt demnach per definitionem auch in allen Teilmengen der Population, also in allen Stichproben, die der Population entnommen werden, seien sie nun repräsentativ oder nicht.

Da eine Selektion lediglich auf die Verteilungsform, also auf die Auftretenswahrscheinlichkeit der Realisationen von Variablen, abzielt, ist der funktionale Zusammenhang, gegeben durch obiges Gleichungssystem, von Selektionen also nicht betroffen. In diesem Sinne ist natürlich die Ladungsmatrix A_{nk} der vorliegenden Faktorenstruktur eine "Invariante" gegenüber Selektionsvorgängen.

Da eine sog. Einfachstruktur nichts weiter als ein spezielles Faktorenmuster ist, gilt das insbesondere auch für Einfachstrukturen.

Veränderung der Kovarianz zwischen Faktoren.

Gegeben sei wieder eine n-dimensionale zufällige Variable, dazu eine Faktorenstruktur (A_{nk}, F, V_{kk}) von X. Ferner sei eine Selektion a bezüglich (X, F) betrachtet (Selektionsvariable seien nicht eigens spezifiziert).

Diese Selektion verändert im allgemeinen die Verteilungsdichte von (X, F) so daß die Kovarianzmatrix $V_{kk} = F' o F$ von F übergeht in V_{kk}^* (vgl. dazu S.81).

Somit lautet die Faktorenstruktur von X nach Selektion (A_{nk}, F, V_{kk}^*) . Selektion verändert also die gegebene Faktorenstruktur einer Variablen insoweit, als die Abhängigkeitsverhältnisse zwischen den Faktoren sich ändern. Waren insbesondere die Faktoren $F_1, \dots, F_k$ vor der Selektion als unabhängig voneinander vorausgesetzt, so wird dies nach der

Selektion i. a. nicht mehr der Fall sein.

Die (vollständige) Kovarianzmatrix von X nach Selektion errechnet sich gemäß (4.1) zu $Q_{nn} = A_{nk} V_{kk}^* A'_{nk}$ unter Berücksichtigung der im vorigen Kapitel ("Konstanz des Faktorenmusters") gemachten Ausführungen.

Kompensatorische Kovarianz im Faktorenmodell.

Der zufälligen Variablen $X = (X_1, \dots, X_n)$ denke man sich die Faktorenstruktur $(A_{nk}, (F_1, \dots, F_k), E)$ mit stochastisch unabhängigen Faktoren $F_1, \dots, F_k$ unterlegt.

Bezeichnet man mit a_{ij} das Element in der i -ten Zeile und der j -ten Spalte von A_{nk} , so gilt nach der Definition (4) der Faktorenstruktur

$$X_i = \sum_{j=1}^k a_{ij} F_j \quad (i = 1, 2, \dots, n).$$

Wie auf Seite 83 Zeile (a) sei X_0 definiert als gewichtete Summe der X_i mit den Gewichten $b_1, b_2, \dots, b_n$:

$$X_0 = \sum_{i=1}^n b_i X_i = \sum_{i=1}^n b_i \left(\sum_{j=1}^k a_{ij} F_j \right) .$$

Bei zweckmäßiger Zusammenfassung der Summanden und Abkürzung entsprechender Faktoren-Koeffizienten durch $a_1, \dots, a_k$ erhält man

$$(24) \quad X_0 = a_1 F_1 + \dots + a_k F_k \quad \text{mit} \quad a_j = \sum_{i=1}^n b_i a_{ij} \quad (j = 1, 2, \dots, k).$$

X_0 kann also auch als gewichtete Summe der Faktoren $F_1, F_2, \dots, F_k$ aufgefaßt werden.

Bei einer Selektion hinsichtlich X_0 , welche die Varianz von X_0 einschränkt, treten also zwischen den Faktoren $F_1, F_2, \dots, F_k$ kompensatorische Kovarianzen auf (s. (19.2)), so daß die Faktoren ihre Unabhängigkeit verlieren. Die Korrelationskoeffizienten zwischen den Faktoren werden z. B. sämtlich negativ, wenn die Ladungssummen auf den einzelnen Faktoren und die Gewichte $b_1, \dots, b_n$ positiv

sind. Kompensatorische Kovarianz zwischen manifesten Variablen $X_1, \dots, X_n$ ist damit über die Grundgleichung rückführbar auf kompensatorische Kovarianz zwischen Faktoren.

Selektion und "faktorielle Invarianz".

In einer Faktorenstruktur FS von unselegierten beobachtbaren Variablen seien sämtliche spezifischen Faktoren untereinander und von allen gemeinsamen Faktoren unabhängig. Dann führt eine Faktorenanalyse insofern zu brauchbaren Ergebnissen, als Matrizen gefunden werden können, die erstens die reduzierte Kovarianzmatrix der Variablen - definiert bezüglich der Faktorenstruktur FS - reproduzieren und zweitens über lineare Transformationen mit der Matrix der Ladungen auf den gemeinsamen Faktoren von FS verknüpft sind. Gesetzt, dies sei der Fall, dann lautet die für die Beurteilung von Selektionseffekten entscheidende Frage:

Wird durch eine Faktorenanalyse von selegiertem Material diese Ladungsmatrix tatsächlich auch zugänglich?

Kann also ein etwa bestehender funktionaler Zusammenhang zwischen beobachtbaren Variablen durch eine Faktorenanalyse auch bei Vorliegen von Selektionen wiedergefunden werden?

Diese Frage läßt sich auf die beiden folgenden, etwas präziser gefaßten Fragen reduzieren:

- a. Existieren zu der reduzierten Kovarianzmatrix - nun definiert bezüglich der Faktorenstruktur FS^* , die den selegierten Variablen unterliegt und die gegenüber FS durch Selektion verändert ist - ebenfalls reproduzierende Matrizen?
- b. Wenn ja, in welcher Beziehung stehen diese reproduzierenden Matrizen zu den reproduzierenden Matrizen hinsichtlich des unselegierten Materials?

Die Antwort auf diese Fragen hängt davon ab, ob die Unabhängigkeit der spezifischen Faktoren im unselegierten Falle durch die Selektion aufgehoben wird oder nicht. Wenn infolge von Selektion auch nur ein spezifischer Faktor seine Unabhängigkeit von einem der übrigen Faktoren verliert, so erhöht sich i. a. nicht nur sofort der Rang der reduzierten Kovarianzmatrix (s. (7.4)), sondern es wird auch die Existenz einer reproduzierenden Matrix überhaupt infrage gestellt (s. (9.1), (9.2)). Also auch dann, wenn im Idealfall die Faktorenstruktur so beschaffen ist, daß eine Faktorenanalyse im unselegierten Fall zu sinnvollen Ergebnissen führt, ist es bei Vorliegen von Selektion meist schon theoretisch nicht mehr möglich, diese Struktur durch eine Faktorenanalyse auch aufzufinden.

Wie die obigen Fragen beantwortet werden müssen, wenn die erwähnte Unabhängigkeit bei Selektion erhalten bleibt, soll nun untersucht werden.

Zunächst vorbereitend sei bemerkt, daß eine etwa bestehende Unabhängigkeit der spezifischen Faktoren sicherlich nur dann erhalten bleibt, wenn die Selektionsvariablen unabhängig von den spezifischen Faktoren sind.

Die Unabhängigkeit der Selektionsvariablen von den spezifischen Faktoren wird auch von MEREDITH (1964, S. 178) in dem Katalog der Voraussetzungen aufgeführt, unter denen "factorial invariance" zu erwarten ist. Bemerkenswert ist allerdings, daß MEREDITH in seinen Ausführungen, also insbesondere auch bei der Herleitung seiner Invarianz-Behauptungen, auf diese Voraussetzung an keiner einzigen Stelle mehr zurückgreift.

Wie sich zeigen wird, ist diese Unabhängigkeit der Selektionsvariablen von den spezifischen Faktoren auch die einzige Forderung, die man von der Theorie her stellen muß, wenn eine praktisch durchgeführte Faktorenanalyse bei Vorliegen von Selektion ebenfalls zu sinnvollen Ergebnissen führen soll.

Diese letzten Überlegungen seien nun näher ausgeführt.

Betrachtet sei eine n -dimensionale zufällige Variable X und eine Faktorenstruktur $FS =: ((A_{np}, B_{nq}), (G, H), V_{kk})$ ($k = p + q$) von X , wobei

G ein p -dimensionaler gemeinsamer Faktor und
 H ein q -dimensionaler spezifischer Faktor sein soll.

$\bar{Q}_{nn}$ bezeichne die reduzierte Kovarianzmatrix von X , bezogen auf die obige Faktorenstruktur FS .

Betrachtet sei ferner eine weitere ein- oder mehrdimensionale zufällige Variable Z . Ob Z und X Komponenten gemeinsam haben, kann offen bleiben.

Schließlich sei eine nicht ausgeartete Selektion a bezüglich $U =: (Z, X, G, H)$ gegeben mit Z als Selektionsvariable bezüglich a , U .

Infolge der Selektion geht die Faktorenstruktur FS über in die Faktorenstruktur FS^* von X . Man erhält FS^* , indem man in FS den Ausdruck V_{kk} formal durch V_{kk}^* ersetzt.

$\bar{Q}_{nn}^*$ bezeichne die reduzierte Kovarianzmatrix von X , nun bezogen auf die Faktorenstruktur FS^* .

Voraussetzungen:

1. H sei stochastisch unabhängig von G und
2. sämtliche Komponenten von H seien untereinander stochastisch unabhängig.

(a) Diese beiden Voraussetzungen sollen sicherstellen, daß es Matrizen gibt, die $\bar{Q}_{nn}$ reproduzieren, unter denen sich A_{np} befindet (s. Satz (9.2)).

3. Z sei stochastisch unabhängig von H .

Mit Hilfe dieser Formulierungen und Voraussetzungen gilt nun der folgende

(25) Reproduktions-Invarianz-Satz:

Behauptungen:

1. A_{np} reproduziert sowohl $\bar{Q}_{nn}$ als auch $\bar{Q}_{nn}^*$.
2. Die Menge der Matrizen, die $\bar{Q}_{nn}^*$ reproduzieren, stimmt mit der Menge derjenigen Matrizen, die $\bar{Q}_{nn}$ reproduzieren, überein.

Beweis:

Da Z nach Voraussetzung 3 unabhängig von H ist, folgt mit Satz (18.1), daß die Dichte $f(h)$ von H sich unter der Selektion nicht verändert, insbesondere bleibt also auch die Unabhängigkeit der Komponenten von H bestehen.

Da Z ferner Selektionsvariable bezüglich $(a, (Z, X, G, H))$ ist, ist nach Satz (16) Z Selektionsvariable auch bezüglich $(a, (Z, G, H))$, also ist nach Satz (17) auch (Z, G) Selektionsvariable bezüglich $(a, (Z, G, H))$. Z, G sind nach den Voraussetzungen 1 und 3 beide unabhängig von H. Nach Satz (18.2) besteht diese Unabhängigkeit dann auch nach der Selektion, insbesondere also bleibt H unabhängig von G.

Die Faktorenstruktur $FS^* = ((A_{np}, B_{nq}), (G, H), V_{kk}^*)$ von X nach Selektion erfüllt also die Unabhängigkeitsvoraussetzungen von Satz (9.2) ebenfalls, also reproduziert A_{np} auch $\bar{Q}_{nn}^*$.

Alle Matrizen, die sowohl $\bar{Q}_{nn}^*$ als auch $\bar{Q}_{nn}$ (s. (a)) reproduzieren, erhält man also durch Anwendung linearer regulärer Transformationen auf A_{np} (s. (12)).

Somit sind die Behauptungen bewiesen.

Mit der Definition (8) der reproduzierenden Matrix ergibt sich aus Satz (25) noch die

(25.1) Folgerung:

$\bar{Q}_{nn}$ und $\bar{Q}_{nn}^*$ haben denselben Rang p.

Wenn die spezifischen Faktoren von den Selektionsvariablen unabhängig sind, ist es also belanglos, ob man selektiertes oder unselektiertes Material einer Faktorenanalyse unterwirft, denn man erhält ja in jedem Falle Zugang zu derselben Menge von reproduzierenden Matrizen.

D. Zusammenfassende Darstellung der bisherigen Ergebnisse.

Im Rückblick auf die theoretischen Ausführungen des zweiten Teils der vorliegenden Arbeit kann man speziell im Hinblick auf Selektionen zweierlei feststellen:

1. Die Matrix der Ladungen auf den Faktoren der Ausgangsfaktorenstruktur, also das Faktorenmuster, bleibt unter Selektion trivialerweise erhalten.

2. Setzt man voraus, daß in der Ausgangsfaktorenstruktur die spezifischen Faktoren untereinander und von den gemeinsamen Faktoren unabhängig sind und ferner die Selektionsvariablen unabhängig von den spezifischen Faktoren sind, so behält auch unter Selektion die Matrix der Ladungen auf den gemeinsamen Faktoren ihre Reproduktions-eigenschaft bei.

Zum Beweis vergleichbarer Aussagen benutzte MEREDITH neben den eben erwähnten Unabhängigkeitsvoraussetzungen noch die Selektionsformeln von LAWLEY mitsamt den hier erforderlichen Voraussetzungen, nämlich der Regularität der Kovarianzmatrix der Selektionsvariablen sowie der Homoskedastizität und Linearität der Regressionen von Nichtselektions- auf Selektionsvariable, und schloß darüber hinaus noch die Selektionsvariablen von den zu analysierenden beobachteten Variablen aus. Dagegen stellt sich hier heraus, daß man von diesen Voraussetzungen lediglich die unter Punkt 2 genannten Unabhängigkeitsvoraussetzungen benötigt.

Bemerkenswert in diesem Zusammenhang ist, daß MEREDITH die LAWLEYSchen Selektionsformeln als für die Theorie der Selektion besonders bedeutungsvoll hinstellt und auch auf die beiden Voraussetzungen der Linearität und Homoskedastizität der erwähnten Regressionen näher eingeht; von der Voraussetzung der Unabhängigkeit zwischen spezifischen Faktoren und den Selektionsvariablen jedoch sagt er ledig-

lich, sie schlieÙe die Möglichkeit einer direkten Selektion hinsichtlich der zu analysierenden beobachteten Variablen aus (s. Fußnote S.19), eine sicherlich irreführende Kommentierung. Da diese Bedingung bei MEREDITH im weiteren dann weder benutzt wird noch sonst in irgendeiner Form zur Sprache kommt, kann sie sogar vom Leser leicht übersehen werden. So zählt z. B. VUKOVICH (1967, S. 117) die wichtigsten MEREDITHschen Voraussetzungen für faktorielle Invarianz insofern unvollständig auf, als er gerade diese Unabhängigkeitsvoraussetzung nicht erwähnt.

Im übrigen aber haben die theoretischen Überlegungen vorwiegend zu Resultaten geführt, die für die Faktorenanalyse negativ sind:

1. Sind bei einem Experiment zwar die Kovarianzen zwischen den beobachteten Variablen, aber nicht die Kommunalitäten empirisch zugänglich - d. h. sind die Werte in der Hauptdiagonale der Kovarianzmatrix nicht bekannt - so kann eine Faktorenanalyse zu keinen sinnvollen Resultaten führen; denn die reproduzierenden Matrizen sind im Hinblick auf die Reproduktion der Werte in der Hauptdiagonale nicht eindeutig bestimmt. Da durch die Kovarianzen allein die Diagonalwerte nicht eindeutig festgelegt sind, hat es auch keinen Zweck, die Kommunalitäten aufgrund dieser Kovarianzen schätzen zu wollen. Bei einer reproduzierenden Matrix, die man trotz ursprünglich fehlender Kommunalitäten erstellt hat, kann man demnach nicht angeben, in welcher Beziehung diese Matrix zu den Parametern des Modells steht, das man den Daten unterlegt hat. So verstanden ist ein auf diese Weise gewonnenes Ergebnis "sinnlos".

2. Aber auch dann, wenn die zu analysierenden Daten durch ein lineares Modell gedeckt werden und neben den Kovarianzen auch die Kommunalitäten bekannt sind, ist nicht sicher, ob eine Faktorenanalyse der reduzierten Kovarianzmatrix zu reproduzierenden Matrizen führen kann; denn hierzu muß die reduzierte Kovarianzmatrix positiv

semi-definit sein, was i. a. nur der Fall ist, wenn die spezifischen Faktoren der unterliegenden Faktorenstruktur untereinander und von den gemeinsamen Faktoren unabhängig sind. Aber selbst wenn trotz Verletzung dieser Unabhängigkeitsbedingung die Positiv-Semi-Definitheit der reduzierten Kovarianzmatrix noch gegeben sein sollte, hätte man immer noch damit zu rechnen, daß deren Rang größer als die Zahl der gemeinsamen Faktoren der Struktur ist.

Für eine Abhängigkeit der spezifischen Faktoren untereinander und von den gemeinsamen Faktoren sprechen mindestens zwei Gründe:

a. Läßt man Abhängigkeit zwischen gemeinsamen Faktoren zu, muß man logischerweise auch abhängige spezifische Faktoren zulassen.

b. Nicht zu vermeidende Selektionsvorgänge führen in der Regel zu abhängigen spezifischen Faktoren, und zwar auch dann, wenn diese in einer fiktiven Ausgangspopulation unabhängig sein sollten; es sei denn, die spezifischen Faktoren seien von den Selektionsvariablen unabhängig. Insofern aber gemeinsame Faktoren von den Selektionsvariablen abhängig sein können, besteht kein rationaler Grund, dasselbe für spezifische Faktoren auszuschließen.

3. Sind alle Bedingungen für die Anwendbarkeit der Faktorenanalyse gegeben, so ist dennoch nicht gewährleistet, daß eine vorliegende Reproduktionsmatrix auch sachlich sinnvoll ist: Es gibt ja viele Matrizen, die eine gegebene reduzierte Kovarianzmatrix reproduzieren; nicht allen jedoch kann man auch Faktoren unterlegen, so daß die Daten gedeckt werden. Einer Reproduktionsmatrix selbst aber kann man nicht ansehen, ob sie in diesem Sinne "sinnvoll" ist oder nicht. Hiermit soll auch gesagt werden, daß die Rotation der Ergebnisse einer Faktorenanalyse problematisch ist. Der im Kapitel "Reproduzierende Matrix und unterlegte Faktoren" exakt belegte Sachverhalt gilt sicherlich in gleicher Weise auch in anderen Fällen, wenn z. B.

zu analysierende Tests deutlich nicht normalverteilt sind infolge von Selektionen oder auch aus anderen Gründen. Nach Durchführung einer Faktorenanalyse mit anschließender orthogonaler Rotation muß man den Ergebnissen stochastisch unabhängige Faktoren unterlegen. Das bedeutet, daß eine Summe von stochastisch unabhängigen Faktoren eine Verteilung haben muß, die von einer Normalverteilung abweicht. Bedenkt man aber, daß die Verteilung einer Summe von stochastisch unabhängigen Variablen einer Normalverteilung immer ähnlicher wird, je mehr Variablen an der Summe beteiligt sind, wird die Problematik zumindest orthogonaler Transformationen deutlich sichtbar.

Es zeigt sich also, daß die Faktorenanalyse - wenigstens in ihrer bisherigen Form - schon von ihren Grundlagen her sich nicht zu einer allgemein brauchbaren Methode entwickeln läßt. Über die Hintergründe hierfür werden im folgenden Kapitel u. a. einige Überlegungen angestellt.

E. Diskussion

1. In den bisherigen Überlegungen wurde stets die reduzierte Kovarianzmatrix der beobachteten Variablen als Ausgangsmatrix für eine Faktorenanalyse angenommen. Man könnte nun auf den Gedanken kommen, die hierdurch entstehenden Schwierigkeiten in einfachster Weise so zu umgehen, daß man nicht die Kommunalitäten, sondern die Gesamtvarianzen der zu analysierenden Variablen in die Hauptdiagonale der Kovarianzmatrix einsetzt. Man weiß zwar, daß man in diesem Falle theoretisch zu ebenso vielen gemeinsamen Faktoren kommt, wie man Ausgangsvariablen hat; aber man braucht ja diejenigen Faktoren, die keine nennenswerten Beiträge zur Gesamtvarianz liefern, nicht zu ak-

zeptieren und nur die restlichen mit dem größten Varianzanteil in der üblichen Weise in interpretierbare Faktorenlösungen zu transformieren.

So bestechend diese Überlegungen zunächst auch erscheinen mögen, auch sie erweisen sich als unhaltbar. Sofern man nämlich den Daten einen Meßfehler zuschreiben muß, ist man gezwungen, ihnen eine Faktorenstruktur mit gemeinsamen und spezifischen Faktoren zu unterlegen; denn nach der herkömmlichen Fehlerdefinition müssen Fehlerfaktoren spezifisch sein. Ladungen auf gemeinsamen Faktoren werden aber nur dann über eine Faktorenanalyse zugänglich, wenn man in die Hauptdiagonale der zu analysierenden Kovarianzmatrix die Kommunalitäten einsetzt. Das Einsetzen der Gesamtvarianz der Variablen ist also in Wirklichkeit ebenfalls eine Kommunalitätenschätzung, nur daß jetzt deren größtmöglichen Werte eingesetzt werden, was in den meisten Fällen einer offensichtlichen Fehlschätzung gleichkommt.

Eine noch bleibende Möglichkeit wäre, die Zuverlässigkeit bzw. die zuverlässige Varianz der Testvariablen als Kommunalitätsschätzungen anzusetzen. Aber, so kann man sich nun fragen, wenn man schon spezifische Faktoren zulassen gewillt ist, wieso sollen ausgerechnet die Fehlerfaktoren die einzigen spezifischen Faktoren sein, die vorkommen können?

2. Allerdings wird hier eine Möglichkeit sichtbar, in Einzelfällen Fragestellungen mit Hilfe einer Faktorenanalyse doch noch einwandfrei bearbeiten zu können: Sie besteht in der Anwendung einer Testbatterie, die so sorgfältig konstruiert worden ist, daß die gesamte zuverlässige Varianz gemeinsame Varianz ist. In diesem Falle hätte man als einzige spezifische Faktoren die Fehlerfaktoren. Herkömmlicherweise werden nun Fehlervariable als solche Variable definiert, die von allen übrigen in Betracht kommenden Variablen unabhängig sind. Jetzt also hätte man drei für die Rechnung einer Faktorenanalyse erforderliche

Bedingungen erfüllt: Erstens wären die spezifischen Faktoren untereinander und von den gemeinsamen Faktoren unabhängig, zweitens müßten auch etwaige Selektionsvariablen von diesen spezifischen Fehlerfaktoren unabhängig sein und drittens hätte man mit den Testzuverlässigkeiten auch die Kommunalitäten vorliegen. Allerdings müßte eine solche Testbatterie erst eigens entwickelt werden; denn die in der Diagnostik verwendeten hochzuverlässigen Tests sind mit Absicht so konstruiert, daß die Untertests möglichst gering miteinander korrelieren, die spezifischen Varianzanteile also maximal sind. Aus verständlichen Gründen nämlich hat man in der Praxis kein Interesse daran, nach zeitraubender Durchführung einer Batterie von Tests im wesentlichen nur einen Aspekt eines komplexen Merkmals erfaßt zu haben.

So verlockend aber die Möglichkeit auch erscheint, durch geschickte Konstruktion zu Forschungszwecken eine Testbatterie zu schaffen, deren gesamtezuverlässige Varianz gemeinsame Varianz ist, so wenig erfolgversprechend scheint dieser Weg tatsächlich zu sein: Selbst bei Analysen von 50 und mehr Intelligenztests erreichen die Kommunalitäten bei weitem nicht die Zuverlässigkeiten der verwendeten Tests. Als Beleg diene eine Untersuchung von HOEPFNER, NIHIRA & GUILFORD (1966), in der einmal 50, einmal 41 Testverfahren faktorenanalytisch bearbeitet wurden. Das arithmetische Mittel der mitgeteilten Testzuverlässigkeiten beträgt 0,74 (bzw. 0,65), dagegen liegt der Durchschnitt der Kommunalitäten bei 0,59 (bzw. 0,54). Bei der Berechnung dieser Durchschnitte wurden 12 (bzw. 14) Tests nicht berücksichtigt, da deren Kommunalitäten z. T. beträchtlich größer (!) als die zugehörigen Reliabilitäten waren. Bezieht man diese Tests jedoch in die Berechnungen mit ein, so erhält man für die durchschnittliche Reliabilität 0,73 (bzw. 0,61) und für die mittlere Kommunalität 0,62 (bzw. 0,57).

Möglicherweise liegt der Grund für diese Diskrepanzen darin, daß selbst in dem für Faktorenanalysen am besten

geeignet erscheinenden Bereich, der Intelligenz nämlich, ein - für alle Probanden in gleicher Weise geltendes - additives Zusammenwirken von Intelligenzfaktoren, wie es in den Grundgleichungen niedergelegt ist, nicht gegeben ist.

3. In der Faktorenanalyse wird beobachteten Merkmalen ein funktionales Modell unterstellt und sodann der Versuch gemacht, die Parameter dieses Modells ausschließlich mit Hilfe von Varianzen und Kovarianzen der Merkmale zu bestimmen. Hierbei stellt sich heraus, daß dieser Versuch nicht gelingen kann, da Varianzen und Kovarianzen allein nicht die dazu benötigten Informationen liefern können. Spezielle Annahmen hinsichtlich der Abhängigkeitsstruktur spezifischer Faktoren, gemacht mit dem Ziel, die benötigten Informationen in das Modell einzubringen, erweisen sich als nicht haltbar. So bleibt als Ausweg, auch die Faktoren direkten Messungen zugänglich zu machen, um so die fehlenden Informationen zu erhalten. Wenn das gelingen sollte, wäre die gesamte Faktorenanalyse als Methode überflüssig geworden; sie wurde ja gerade deswegen erdacht, weil die Faktoren des unterlegten Modells nicht beobachtbar sind. Hier wird das Dilemma der Faktorenanalyse deutlich sichtbar: Es steckt im Ansatz der "latenten Struktur".

Normalerweise geht man bei der Modellbildung so vor, daß man zwischen den unabhängigen Variablen (den Eingangsgrößen des Modells) und den abhängigen Variablen (den Ausgangsgrößen des Modells) eine funktionale Beziehung annimmt und durch systematische Variation der unabhängigen Variablen die Parameter des Modells zu bestimmen sucht. Anschließend versucht man, mit Hilfe des auf diese Weise gewonnenen Zahlenmaterials quantitative Vorhersagen über den Ausfall anderer Experimente zu machen. Der Vergleich dieser Vorhersagen mit den tatsächlichen Beobachtungen entscheidet, ob die Modellannahmen zutreffend waren oder nicht.

Bei der Faktorenanalyse scheint man in ähnlicher Weise vorzugehen: Die Faktoren werden als unabhängige Variable aufgefaßt, die beobachteten Merkmale als abhängige Variable, und der unterstellte funktionale Zusammenhang zwischen beiden Sorten von Variablen ist niedergelegt in der Grundgleichung (s. dazu auch BURT, 1958, S. 123). Hier aber hört die Analogie bereits auf: Die Faktoren des Modells sind der Messung nämlich nicht zugänglich. Somit entpuppen sich die Grundgleichungen als (operationale) Definitionsgleichungen für die Faktoren.

Solche Definitionen neuer Größen müssen - wenigstens im Prinzip - die (eindeutige) Erfassung oder "Messung" der definierten Größe ermöglichen. Auf die Faktorenanalyse bezogen heißt das: Ausgehend von den Testwerten eines Probanden muß dessen Faktorenausstattung bestimmbar sein. Inwieweit das bei der Faktorenanalyse der Fall ist, soll hier nicht weiter untersucht werden, es sei lediglich darauf hingewiesen, daß bei HARMAN (1960, S. 337-361) mehrere Verfahren zur Schätzung von Faktorenwerten beschrieben werden. Allerdings werden zu Forschungszwecken solche Faktorenmessungen nur selten auch durchgeführt.

Aber selbst wenn solche Faktorenmessungen tatsächlich durchgeführt werden, von wesentlichem theoretischen Nutzen sind sie kaum. Denn die Grundgleichungen, über die die Faktoren definiert sind, stellen auch die einzige formelmäßige Verknüpfung dar, in der Faktoren mit anderen Größen, seien es nun Testvariable oder andere Faktoren, verknüpft sind. Ein Analogon hierzu wäre es, wenn in der Physik eine Größe β definiert würde durch

$$\beta =: v - 2b + 3m^2 - 4h + 5V - 6r,$$

wobei v , b , m , h , V , r meßbare Größen bedeuten sollen; die so definierte Größe β aber, außer in dieser Definitionsgleichung, in keiner anderen Gleichung mehr vorkommt.

Man kann β in diesem Falle zwar messen; aber es kann nicht entschieden werden, ob β eine zweckmäßig definierte Größe ist oder keinerlei theoretische und praktische Bedeutung hat.

Wiederum auf die Faktorenanalyse bezogen heißt das: Solange Faktorenwerte lediglich "gemessen", aber nicht in anderen Zusammenhängen als unabhängige Variable auftreten, kann sich ein faktorieller Ansatz weder als brauchbar noch als unbrauchbar erweisen. Das ist auch der Grund dafür, warum faktorenanalytisch gewonnene Ergebnisse stets diese "erschreckende Unverbindlichkeit" zeigen, auf die ORLIK erst vor kurzem noch hingewiesen hat (ORLIK, 1967, S. 87).

4. Bleibt man dabei, die Grundgleichung nicht als Modell, sondern als Definition und Ausgangspunkt für die Messung der Faktoren aufzufassen, so sind die Ladungen auf den so definierten Faktoren nichts weiter als Proportionalitäts-Koeffizienten, analog den Zahlen 1, -2, 3, -4, 5 und -6 in der Definitionsgleichung der fiktiven Größe β . Die zahlenmäßige Bestimmung dieser Koeffizienten mit Hilfe einer Faktorenanalyse ist demnach ein Analogon zur Eichung eines Meßinstruments. So gesehen sind die zahlreichen Veröffentlichungen der Ergebnisse von Faktorenanalysen tatsächlich nur Mitteilungen von Eich-Koeffizienten ganz spezieller Meßinstrumente, nämlich der in einer bestimmten Situation an bestimmten Probanden angewendeten Testbatterie, wobei jedoch nach erfolgter Eichung auf die Durchführung der Messungen selbst verzichtet wurde, ein in den Naturwissenschaften sicherlich nicht häufiges Vorgehen. Die Verwechslung von Eichung und Messung im Terrain der Faktorenanalyse hat ihren Grund in folgendem: Wegen des beträchtlichen mathematischen Aufwandes wird von vielen Anwendern der Faktorenanalyse möglicherweise nicht klar erkannt, daß die Grundgleichungen, operational gesehen, kein Modell darstellen, sondern die Faktoren erst definieren.

An dieser Stelle sei bemerkt, daß auch noch in anderen Gebieten der Psychologie möglicherweise operationale Definitionen für Modelle gehalten werden; man denke nur an Fälle von ein- und mehrdimensionalen Skalierungen oder an Zerlegungen von Lernkurven mit Hilfe eines von TUCKER (1960) entwickelten Verfahrens.

5. Verfechter der Faktorenanalyse werden nun sicherlich anführen, daß sie keineswegs einem solchen Realismus huldigten, die durch eine Faktorenanalyse zutage geförderten Faktoren nun auch wirklich als "Naturgrößen" anzusehen. Vielmehr fasse man die gefundenen Faktoren als zunächst fiktive Dimensionen unabhängiger Variabilität auf, sodann interessiere man sich dafür, welchen Anteil die einzelnen manifesten Merkmale an der Variation in den einzelnen Dimensionen haben. Ich muß gestehen, daß ich solche Charakterisierungen der Faktorenanalyse nie verstanden habe. Daß eine positiv semi-definite Matrix, um die es sich bei einer Kovarianzmatrix mit passend ausgefüllter Hauptdiagonale handelt, als Produkt einer Matrix mit ihrer Transponierten darstellbar ist, ist eine mathematisch erwiesene Identität und hat mit der Art und Weise, wie diese Zahlen der Ausgangsmatrix zustandegekommen sind, überhaupt nichts zu tun. Daß ferner fast jede Funktion innerhalb gewisser Grenzen linear (z. B. durch eine Gerade) angenähert werden kann, ist ebenfalls eine bekannte Tatsache. Jedem Satz von beobachteten Variablen kann also in erster Annäherung eine lineare "latente Struktur" unterstellt werden, gleichgültig, auf welche Weise die Variablen tatsächlich strukturiert sind. Somit kann eine faktorenanalytische Untersuchung wissenschaftlich nur dann sinnvoll sein, wenn der lineare Ansatz aus der Sache begründbar ist.

Die Anwendbarkeit der Mathematik in einer Sachwissenschaft beruht darauf, daß erstens gewissen Objekten der Sachwissenschaft bestimmte mathematische Gebilde zugeordnet werden können, zweitens daß Beziehungen zwischen

den Sachobjekten definiert werden können und drittens diese Beziehungen anschließend in Beziehungen zwischen den mathematischen Gebilden übersetzt werden können. In diesem Falle besteht zwischen der Sache und ihrer mathematischen Beschreibung eine (teilweise) Isomorphie. Nur insofern diese besteht, besteht auch zwischen den durchgeführten mathematischen Operationen und den sachlichen Gegebenheiten eine - mehr oder weniger vollständige - Entsprechung, nur dann also bildet die Mathematik die intendierten sachlichen Gegebenheiten auch ab. Hierher stammt auch der Name "Modell" für die mathematische Entsprechung. Sofern ein solcher Bild-Urbild-Zusammenhang jedoch nicht gegeben ist, sind durchgeführte Rechnungen nichts weiter als einzelne zahlenmäßige Bestätigungen von mathematisch in allgemeiner Form bereits bewiesenen Tatbeständen. Welchen Sinn soll die Faktorenanalyse haben, wenn man darauf verzichtet, sie als Modell "für etwas" anzusehen? Hierbei räume ich allerdings ein, daß ich Varianzen und Kovarianzen nicht als Modelle psychologisch relevanter Sachverhalte ansehen, sondern als Maße von Variationen innerhalb der Ausprägungsgrade psychologischer Merkmale.

6. Befürworter charakterisieren die Faktorenanalyse u. a. als Methode zur Kategorisierung bzw. ökonomischeren Beschreibung von Daten (s. z. B. FISENCK 1953; HARMAN 1960, S. 5), wobei zuweilen auch Vergleiche gezogen werden zwischen den Ergebnissen von Faktorenanalysen und etwa dem LINNischen System in der Botanik, dem periodischen System der Elemente in der Chemie oder anderen Klassifikationsschemata.

Es soll daher im folgenden diskutiert werden, inwieweit diese Kennzeichnung als "Datenreduktions-Methode" auf die Faktorenanalyse zutrifft.

a. Die Faktorenanalyse als Instrument zur Klassifizierung von Tests.

In der Faktorenanalyse werden Tests unter einem

verallgemeinerten Gültigkeitsaspekt betrachtet; denn es wird die Varianz von Tests aufgespalten in einzelne Anteile, die man auf die Varianz (unabhängiger) Faktoren zurückführt. An den Ladungen eines Tests auf solchen Faktoren müßte man demnach erkennen können, in welchem Ausmaß der Test was mißt, und zwar im Vergleich zu dem, was die übrigen Tests der Batterie messen.

Die Kennzeichnung eines Tests hinsichtlich dessen, was er erfaßt, geschieht also durch Angabe seiner Ladungen auf den Faktoren, die somit den Kategorien eines mehrdimensionalen Ordnungsschemas entsprechen.

Ökonomisch wäre diese Art der Kategorisierung von Tests in dem Augenblick, wo die Zahl der Kategorien geringer wäre als die Zahl der in die Analyse eingegangenen Tests.

Nun beträgt aber die Bedeutsamkeitsgrenze, unterhalb derer man Ladungen auf gemeinsamen Faktoren gewöhnlich nicht mehr beachtet, etwa 0,3, was einem relativen Varianzanteil von rund 0,1 oder 10 % entspricht. Die Ladungen auf den testspezifischen Faktoren liegen aber meistens beträchtlich über dieser Bedeutsamkeitsgrenze, so daß man die spezifischen Faktoren notgedrungen zur Kategorisierung der Tests mitverwenden muß. Somit braucht man in diesen Fällen zur Beschreibung der Tests nicht weniger, sondern sogar mehr Kategorien als Tests vorliegen!

Natürlich kann man auf eine solche inhaltlich vollständige Kategorisierung von Tests verzichten und die Kategorisierung lediglich hinsichtlich der gemeinsamen Faktoren vornehmen; hinsichtlich dessen, was sie erfassen, hat man die Tests nun allerdings nicht mehr klassifiziert, insofern man wichtige Ordnungsgesichtspunkte willkürlich außer Acht gelassen hat. Ein so gewonnenes Ordnungsschema ist demnach ebenso willkürlich wie andere gleichdimensionale Schemata auch, eine Faktorenanalyse wäre in diesem Falle also eigentlich überflüssig gewesen.

An dieser Stelle wird auch sichtbar, daß der Vergleich faktorenanalytischer Ergebnisse mit den genannten beiden

Ordnungssystemen in Botanik und Chemie nicht nur hinkt, sondern tatsächlich völlig unangebracht ist: Letztere Systeme ordnen Objekte der betreffenden Wissenschaften; die Faktorenanalyse hingegen ordnet primär nicht Gegenstände der Psychologie (z. B. "Fähigkeiten"), sondern "Meßinstrumente", nämlich die Tests, mit denen die psychischen Gegebenheiten erfaßt werden sollen.

b. Die Faktorenanalyse als Instrument zur Klassifizierung von Probanden.

Als zweites, was durch eine Faktorenanalyse ökonomischer beschrieben werden könnte, kämen getestete Probanden infrage, sofern statt vieler einzelner Testergebnisse nur einige wenige Faktorenwerte angegeben zu werden brauchten, um einen bestimmten Probanden zu kennzeichnen. De facto aber erreicht man auch hier die "Ökonomie" erst dadurch, daß man die Ausstattung der Probanden hinsichtlich der spezifischen Faktoren willkürlich unberücksichtigt läßt. Im Prinzip also treffen hier die gleichen Argumente zu wie vorhin unter Diskussionsbeitrag 6. a..

c. Die Faktorenanalyse als Instrument zur Klassifizierung von "Eigenschaften".

Insbesondere gegen den letzten Absatz unter Punkt 6. a. könnte man einwenden, daß die Faktorenanalyse schließlich die Aufstellung fundierter Intelligenz- und Persönlichkeitsmodelle erst ermöglicht habe. Diese Argumentation läßt sich aber sicherlich nicht aufrechterhalten. Wie aus den Diskussionsbeiträgen 4 und 5 zu entnehmen ist, impliziert faktorenanalytisches Vorgehen die operationale Definition der zu erfassenden Gegenstände, wobei jedoch diese Definition auch die einzige Meßmöglichkeit der so definierten Größen liefert. In einem solchen Falle - so ist meine Auffassung - kann man fast jedes Klassifikationsmodell, das man a priori an die Daten heranträgt, in irgendeiner Weise faktorenanalytisch rechtfertigen - es muß

nur in sich logisch konsistent und sachlich gesehen "vernünftig" sein; denn es ist ja kaum eine Möglichkeit für eine Überprüfung vorhanden. Die Tatsache der Existenz mehrerer unterschiedlicher Intelligenz- und Persönlichkeitsmodelle, die alle faktorenanalytisch untermauert sind, ist sicherlich eine Stütze dieser Auffassung.

Ein hierzu analoges Vorgehen in den Naturwissenschaften ist nicht leicht aufzuweisen, da die Faktorenanalytiker in dieser Hinsicht "einmalig" vorgehen, es sei denn, man ginge bis auf Ptolemäus zurück. Ein vergleichbarer Fall läge in etwa vor, wenn es in der Chemie nur eine einzige Meßmöglichkeit für diejenigen chemischen Eigenschaften gäbe, hinsichtlich derer die Elemente im periodischen System angeordnet werden sollen - nämlich für Atomgewicht und Wertigkeit - und Atomgewicht und Wertigkeit gleichzeitig durch diese Meßvorschrift erst definiert würden.

Somit wird einsichtig, daß die Faktorenanalyse, praktisch gesehen, entweder nicht ökonomischer klassifiziert oder aber keine Klassifikationsgesichtspunkte liefert, die vor anderen sachlich besonders ausgezeichnet sind.

7. Schließlich sei noch auf Faktorenanalysen eingegangen, die in Untersuchungen zur Erfassung von Gruppenunterschieden gerechnet worden sind. Bei solchen Untersuchungen muß man naturgemäß auf Gruppen von Individuen zurückgreifen, die hinsichtlich eines oder auch mehrerer Merkmale relativ homogen sind, z. B. hinsichtlich des Alters, der Intelligenz oder der Leistung. Bezogen auf die Gesamtpopulation handelt es sich hier um hoch ausgelesene Stichproben, so daß mit starken Selektionseffekten zu rechnen ist. Zur Verteidigung solchen Vorgehens kann man anführen, daß man in diesen Untersuchungen als Population die Menge derjenigen Individuen definiere, die sich bezüglich des betreffenden Merkmals eben nicht wesentlich voneinander unterscheiden. Somit ist die untersuchte Stichprobe per definitionem unausgelesen. Gegen diese Art der Definition der Population ist nichts einzuwenden, wohl

aber gegen die Folgerungen, die daraus gezogen werden. Nur bei Erfüllung bestimmter Unabhängigkeitsvoraussetzungen, die Faktoren betreffend, kann eine Faktorenanalyse sinnvoll durchgeführt werden. Ob aber Zufallsvariable, bezogen auf eine definierte Population, untereinander unabhängig sind oder nicht, ist wahrhaftig nicht in das Belieben des Untersuchers gestellt.

Zusammenfassung

Nach Ausarbeitung von Grundzügen einer Theorie der Faktorenanalyse und Selektion stellt sich folgendes heraus:

a. Die Voraussetzungen, bei denen man auch unter Selektion mit gewissen invarianten Ergebnissen von Faktorenanalysen rechnen kann, lauten:

1. Die spezifischen Faktoren sind untereinander und von den gemeinsamen Faktoren vor der Selektion stochastisch unabhängig.
2. Die Unabhängigkeit der spezifischen Faktoren wird durch die Selektion nicht aufgehoben.

b. Invarianzbehauptungen anderer Autoren sind zum Teil unter erheblich mehr Voraussetzungen als in der vorliegenden Arbeit bewiesen worden, zum Teil zeigte sich auch, daß sie per definitionem galten oder daß es sich um sinnlose Aussagen handelte.

c. Unvermeidbare Selektionen oder Vorgänge, die Selektionen analog sind, führen im allgemeinen zu abhängigen Faktoren, gleichgültig, ob es sich dabei um spezifische oder gemeinsame handelt. Das hat zur Folge, daß eine reduzierte Kovarianz- bzw. Korrelationsmatrix im allgemeinen nicht faktorisierbar ist.

Die Faktorenanalyse ist demnach keine allgemein anwendbare Methode. Ob das Konzept der latenten Struktur, so wie es in der Faktorenanalyse und verwandten Verfahren Anwendung findet, in anderen Naturwissenschaften eine Parallele hat, wird darüber hinaus bezweifelt.

A N H A N G

Zahlenbeispiele für Selektionseffekte

Die folgenden drei Beispiele sollen dazu dienen, Selektionseffekte in einfachen Fällen zu veranschaulichen. In den ersten beiden Beispielen wird gezeigt, wie sich Trunkationen von Summenvariablen auf die Korrelationskoeffizienten zwischen den Einzelvariablen zahlenmäßig auswirken. Die hier angewendete Technik ist die "Simulation" statistischer Variablen mit Hilfe von Zufallszahlen. Das letzte Beispiel zeigt, wie Selektionen die Ergebnisse von Faktorenanalysen beeinflussen können. Die Berechnungen hierfür sind "modellmäßig exakt" durchgeführt. Die Rechnungen selbst wurden im Rechenzentrum der Universität Marburg durchgeführt.

1. Beispiel: Trunkation der Summe von zwei unabhängigen zufälligen Variablen.

Ausgangspunkt für statistische Simulationsexperimente ist ein sog. Zufallszahlen-Generator. Führt man solche Experimente mit Hilfe einer elektronischen Rechananlage durch, so handelt es sich bei dem Zufallsgenerator um ein Unterprogramm, das bei jedem Aufruf eine Zufallszahl (besser: eine Pseudo-Zufallszahl, da die im Rechner ablaufenden Vorgänge determiniert sind) liefert, vergleichbar mit dem Ziehen (mit Zurücklegen) von Losnummern nach Zufall aus einer großen Menge von Losen.

Der in den hier beschriebenen Simulationsexperimenten benutzte Generator erzeugt normalverteilte Zufallszahlen mit dem Erwartungswert 0 und der Varianz 1. Ließe man diesen also z. B. 1000 Zufallszahlen erzeugen, so erhielte man als Mittelwert der erzeugten Zahlen einen Wert um 0 und als Varianz einen Wert um 1. Die relativen Häufigkeiten für das Auftreten bestimmter Zahlen würde der Gaußschen Normalverteilungskurve entsprechen.

Ein solcher Zufallszahlen-Generator ist äquivalent einer normalverteilten Zufallsvariablen mit dem Erwartungswert 0 und der Varianz 1; erzeugte Zufallszahlen sind gleichbedeutend mit Realisationen dieser Zufallsvariablen. Multipliziert man jede dieser Realisationen mit einer Konstanten c und addiert eine Konstante a hinzu, so erhält man ersichtlich eine zufällige Variable mit dem Erwartungswert a und der Varianz c^2 .

Man kann mit demselben Zufalls-Zahlengenerator auch zwei und mehr stochastisch unabhängige zufällige Variable simulieren. Dies ist möglich, da die gelieferten Zufallszahlen stochastisch unabhängig sind. Der Generator erzeugt zu diesem Zweck zunächst fortlaufend Zufallszahlen, die in Zahlenpaaren angeordnet werden. Die erste Zahl x_1 jedes der so gebildeten Paare wird als Realisation einer normalverteilten zufälligen Variablen X_1 aufgefaßt und die zweite Zahl x_2 als Realisation einer davon unabhängigen normalverteilten zufälligen Variablen X_2 . Der "wahre" Korrelationskoeffizient zwischen X_1 und X_2 muß demnach gleich 0 sein.

Eine dritte zufällige Variable sei definiert als

$$X_0 =: 1/\sqrt{2}X_1 + 1/\sqrt{2}X_2,$$

sie stellt eine Summenvariable dar.

Zu einer Realisation (x_1, x_2) von (X_1, X_2) erhält man die zugehörige Realisation x_0 von X_0 als

$$x_0 = 1/\sqrt{2} x_1 + 1/\sqrt{2} x_2.$$

Theoretisch ist die Varianz von X_0 gleich 1 und der Erwartungswert gleich 0, während der "wahre" Korrelationskoeffizient zwischen X_0 und X_1 bzw. X_2 $1/\sqrt{2} = 0,707$ beträgt. Als Summe zweier normalverteilter Variablen ist auch X_0 normalverteilt.

In einer ersten Simulation nun wurden 1000 Zahlentripel (x_0, x_1, x_2) auf die beschriebene Weise erzeugt. Die empirische Varianz der so erhaltenen drei Zahlenreihen und die Kovarianzen bzw. Korrelationskoeffizienten zwischen ihnen sind in Tabelle 1a dargestellt; die empirischen Werte stimmen mit den theoretischen, abgesehen von Stichprobenfehlern, überein.

In weiteren Simulationen wurden zur Berechnung der empirischen Varianzen, Kovarianzen und Korrelationskoeffizienten nur solche vom Zufallszahlen-Generator erzeugte Zahlentripel (x_0, x_1, x_2) herangezogen, bei denen x_0 folgender Bedingung genügte:

$$(a) \quad -c \leq x_0 \leq +c$$

Lag x_0 außerhalb der Grenzen $\pm c$, so wurde das betreffende Zahlentripel nicht berücksichtigt. Für c wurden nacheinander die Werte 1,645, 1,282 und 0,6745 eingesetzt und jeweils 1000 der Bedingung (a) entsprechende Zahlentripel erzeugt. Die Werte für c sind so gewählt worden, daß theoretisch die relative Häufigkeit der links und rechts vom Mittelwert 0 abgeschnittenen Extremwerte von X_0 0,1, 0,2 oder 0,5 betrug (s. z-Tabelle), d. h., bezogen auf die insgesamt erzeugten Tripel - einschließlich der nicht berücksichtigten - wurden auf jeder Seite je 5, 10 oder 25 % extreme Ausprägungsgrade von X_0 abgeschnitten. Die Ergebnisse dieser Simulationsexperimente findet man in den Tabellen 1b - 1d. Besonders auffallend sind hier die betragsmäßig hohen negativen Korrelationskoeffizienten zwischen den ursprünglich unabhängigen zufälligen Variablen X_1 und X_2 .

Zur Demonstration sind in Tabelle 1e die mit Hilfe der Selektionsformeln (19) (s. S. 82) ausgerechneten Varianzen,

Kovarianzen und Korrelationskoeffizienten für die letzten Simulationen angegeben. Die Kenntnis, daß infolge eines Selektionsvorganges die Varianz der Selektionsvariablen X_0 auf 0,139 eingeschränkt worden ist, genügt ja zur Berechnung der übrigen Varianzen, Kovarianzen und Korrelationen. Die Übereinstimmung mit den direkt berechneten Werten in Tabelle 1d ist gut; Abweichungen sind auf Stichprobenfehler zurückzuführen.

Tabelle 1: Die bei den Simulationsexperimenten des ersten Beispiel erhaltenen Varianzen, Kovarianzen und Korrelationskoeffizienten. In den Hauptdiagonalen der nachfolgend aufgeführten Matrizen stehen jeweils die Varianzen der Variablen, darüber die Kovarianzen und darunter die Korrelationskoeffizienten zwischen den Variablen.

Tabelle 1a: 1. Simulation.
Keine Selektion.
Erzeugte Zahlentripel: 1000

	X_0	X_1	X_2
X_0	1,019	0,692	0,720
X_1	0,693	0,982	-0,003
X_2	0,749	-0,002	1,061

Tabelle 1b: 2. Simulation.
Variationsbereich: $\pm 1,645$.
Von 1118 erzeugten Zahlen-
tripeln lagen 118 (= 10,5 %)
außerhalb (theoretisch 10 %).

	X_0	X_1	X_2
X_0	0,656	0,455	0,472
X_1	0,601	0,875	-0,231
X_2	0,615	-0,261	0,899

Tabelle 1c: 3. Simulation.
Variationsbereich: $\pm 1,282$.
Von 1240 erzeugten Zahlen-
tripeln lagen 240 (= 19,4 %)
außerhalb (theoretisch 20 %).

	X_0	X_1	X_2
X_0	0,448	0,310	0,324
X_1	0,544	0,726	-0,287
X_2	0,560	-0,390	0,745

Tabelle 1d: 4. Simulation.
Variationsbereich: $\pm 0,6745$.
Von 1996 erzeugten Zahlen-
tripeln lagen 996 (= 49,8 %)
außerhalb (theoretisch 50 %).

	X_0	X_1	X_2
X_0	0,139	0,084	0,113
X_1	0,303	0,554	-0,436
X_2	0,392	-0,758	0,595

Tabelle 1e: Mit Hilfe der LAWLEYSchen Selektions-Formeln berechneten Varianzen, Kovarianzen und Korrelationskoeffizienten der 4. Simulation.

	X_0	X_1	X_2
X_0	0,139	0,099	0,100
X_1	0,347	0,579	-0,434
X_2	0,353	-0,755	0,571

2. Beispiel: Trunkation der Summe von sieben abhängigen zufälligen Variablen.

Um zu zeigen, welche Bedeutung Selektionen für die praktische Forschung haben können, sei die bereits zitierte Arbeit von REINERT et al. (s. S. 6) näher betrachtet. Die Autoren testeten hier zunächst zwei Gruppen G_1 und G_2 von Jugendlichen (Lebensalter 9;7 bis 11;6 bzw. 11;7 bis 13;6) mit dem Leistungsprüfsystem (LPS) (HORN, 1962) und dem Begabungstestsystem (BTS) (HORN, 1956). Zur Parallelisierung beider Gruppen wählen REINERT et al. anschließend sieben Untertests des LPS aus, nämlich die Tests 1, 3, 6, 9, 12, 14, 15, die im weiteren mit den Zahlen 1 bis 7 durchnummeriert werden. Für jeden der getesteten Probanden werden nun die Rohwerte in diesen sieben Tests zunächst auf die Teststreuungen bezogen und sodann addiert. Die Verteilung dieser Summen für die beiden Stichproben G_1 und G_2 kann man einer von REINERT et al. (1965, S. 263) mitgeteilten Tabelle entnehmen. Die Abbildungen 5.1 und 5.2 zeigen diese Verteilungen - genauer: die zugehörigen Häufigkeitsfunktionen - in graphischer Darstellung.

Aus den Stichproben G_1 und G_2 gewinnen REINERT et al. zwei neue Stichproben A_1 und A_2 von solchen Versuchssper-

sonen, bei denen die Summe der sieben Testwerte zwischen die Zahlen $a = 145$ und $b = 185$ fällt (schraffierter Bereich der in den Abbildungen 5.1 und 5.2 dargestellten Häufigkeitsfunktionen). An Hand dieser ausgelesenen Stichproben A_1 und A_2 berechnen REINERT et al. nun die Interkorrelationen zwischen allen Untertests vom LPS und vom BTS. Tabelle 2 zeigt Auszüge der von REINERT et al. (1965, S. 276 und S. 277) angegebenen Korrelationsmatrizen für die beiden Gruppen A_1 und A_2 , die nur die sieben zur Auslese verwendeten Tests enthalten. Zum Vergleich dazu sind in Tabelle 3 die von HORN (1962) in der Handanweisung zum LPS mitgeteilten Interkorrelationskoeffizienten dieser sieben Tests, berechnet an 200 Erwachsenen, aufgeführt. Es fallen in Tabelle 2 die relative Kleinheit der Korrelationskoeffizienten und die große Anzahl negativer Korrelationskoeffizienten auf.

Tabelle 2: Korrelationen zwischen sieben Untertests des LPS in den Gruppen A_1 und A_2 (nach REINERT et al. 1965). Oberhalb der Diagonale stehen die Korrelationen der Gruppe A_1 , darunter diejenigen der Gruppe A_2 .

	1	2	3	4	5	6	7
1		-0,15	0,05	-0,260	0,38	-0,23	-0,05
2	0,02		-0,03	0,05	-0,18	0,05	0,06
3	0,12	-0,20		0,06	0,14	-0,01	-0,08
4	0,03	0,20	-0,15		-0,10	0,16	-0,13
5	0,34	0,01	-0,03	-0,05		-0,16	-0,11
6	-0,33	-0,16	0,04	-0,04	-0,06		0,16
7	-0,02	-0,13	0,00	0,04	0,09	0,08	

Tabelle 3: Korrelationen zwischen sieben Untertests des LPS (nach HORN, 1962).

Oberhalb der Diagonale stehen die originalen Korrelationen (nach HORN, 1962), darunter die über die Faktorenlösung in Tabelle 5 reproduzierten Korrelationen.

	1	2	3	4	5	6	7
1		0,61	0,60	0,43	0,77	0,40	0,54
2	0,56		0,57	0,63	0,61	0,43	0,47
3	0,63	0,59		0,50	0,66	0,41	0,43
4	0,44	0,62	0,51		0,49	0,48	0,38
5	0,77	0,59	0,66	0,47		0,52	0,47
6	0,46	0,47	0,40	0,44	0,48		0,49
7	0,51	0,46	0,42	0,41	0,53	0,49	

Das Vorgehen von REINERT et al. wurde nun nachgeahmt. Hierzu waren die folgenden drei Schritte notwendig:

1. Schritt: Erstellung einer simultanen Verteilung mit vorgegebenen Parametern von sieben zufälligen Variablen.

Dazu wurden die originalen Korrelationskoeffizienten zwischen den sieben Tests des LPS (s. Tabelle 3) einer Faktorenanalyse unterworfen, die drei Faktoren ergab, mit deren Hilfe die ursprüngliche Korrelationsmatrix gut reproduziert werden konnte. Die Matrix A_{73} der Ladungen auf diesen Faktoren ist in Tabelle 5 zu finden, es handelt sich hierbei um die nach dem Varimax-Kriterium (s. HARMAN, 1960, S. 301) orthogonal rotierte Ausgangslösung. Bezeichnet man die drei gemeinsamen Faktoren mit G_1 , G_2 und G_3 , die test-spezifischen Faktoren mit S_1 , S_2 , ..., S_7 , die Elemente von A_{73} mit a_{ij} und die Ladungen auf den spezifischen Faktoren mit b_i , so sind durch

$$X_i =: a_{i1}G_1 + a_{i2}G_2 + a_{i3}G_3 + b_iS_i, \quad i = 1, 2, \dots, 7$$

sieben zufällige Variablen $X_1, \dots, X_7$ so definiert, daß

die "wahren" Korrelationskoeffizienten zwischen ihnen von den Korrelationskoeffizienten der sieben benutzten Untertests des LPS nur unwesentlich abweichen.

Mit Hilfe des im vorigen Beispiel beschriebenen Zufallszahlen-Generators wurden nun "künstliche Probanden" hergestellt: Zunächst wurden drei Zufallszahlen g_1 , g_2 und g_3 erzeugt, die einen "Probanden" hinsichtlich der drei gemeinsamen Faktoren kennzeichneten. Sodann waren noch sieben weitere Zufallszahlen s_1 , ..., s_7 zur Festlegung der Werte des betreffenden "Probanden" hinsichtlich der sieben spezifischen Faktoren erforderlich. Somit ergaben sich die "Testwerte" x_i eines "Probanden" zu

$$x_i = a_{i1}g_1 + a_{i2}g_2 + a_{i3}g_3 + b_i s_i, \quad i = 1, 2, \dots, 7.$$

Bei Herstellung sehr vieler solcher "Probanden" und Berechnung der empirischen Korrelationskoeffizienten zwischen deren "Testnoten" wird man also Werte erhalten, die von den in Tabelle 3 mitgeteilten bis auf zufällige Abweichungen übereinstimmen. *)

2. Schritt: Berechnung des Variationsbereiches, auf den die aus den X_1 , ..., X_7 gebildete Summenvariable eingeschränkt werden soll.

Der Testsumme bei REINERT et al. entspricht hier die Summenvariable

*) Dieses Verfahren zur Herstellung multivariater Verteilungen mit vorgeschriebenen Parametern wurde bereits von FISCHER & ROPPERT (1964) angewendet, wie mir Herr Dr. FISCHER auf eine Anfrage im März 1966 mitgeteilt hat. Unabhängig davon ist dieses Verfahren von mir in einer früheren Arbeit (KALVERAM, 1965) entwickelt und benutzt worden. Schließlich ist in den Psychometrika 30 (1965, 103-113) dasselbe Verfahren von WHERRY et al. breit beschrieben worden.

$$X_0 =: X_1 + X_2 + \dots + X_7;$$

die Summennote einer Versuchsperson bei REINERT et al. ist hier gleichbedeutend der Summe

$$x_0 = x_1 + x_2 + \dots + x_7$$

eines künstlichen "Probanden". Da die Variablen $X_1, \dots, X_7$ sich normalverteilen mit der Varianz 1, besitzt auch X_0 eine Normalverteilung. Die "wahre" Varianz von X_0 berechnet sich zu

$$\sigma^2 = \sum_{i=1}^7 \sum_{j=1}^7 r_{ij},$$

wobei mit r_{ij} ($i \neq j$) der wahre Korrelationskoeffizient zwischen X_i und X_j gemeint ist und r_{ii} gleich 1 gesetzt ist.

Die Korrelationskoeffizienten r_{ij} ($i \neq j$) erhält man durch Ausrechnen des Matrizenproduktes $A_{73}A'_{37}$ (A_{73} s. Tabelle 5). Man erhält auf diese Weise als Streuung von X_0 :

$$\sigma = 5,37.$$

Unterstellt man, daß es sich bei den in den Abbildungen 5.1 und 5.2 dargestellten Häufigkeitskurven der Testsummen tatsächlich um Normalverteilungsfunktionen handelt, kann man mit Hilfe der dort angegebenen Häufigkeiten die Bereichsgrenzen $a = 145$ und $b = 185$ zunächst in z-Werte der Standard-Normalverteilung umrechnen. Man erhält so

$z_a = -0,129$, $z_b = 1,393$ für die erste Gruppe
und $z_a = -1,787$, $z_b = -0,028$ für die zweite Gruppe.

Multiplikation dieser z-Werte mit σ ergibt schließlich die beiden Variationsbereiche für X_0 , die dem Variationsbereich von $a = 145$ bis $b = 185$ der absoluten Testsummen in den Gruppen A_1 und A_2 bei REINERT et al. entsprechen:

Für die erste Gruppe erhält man

$$a' = z_a \cdot \sigma = -0,692, \quad b' = z_b \cdot \sigma = 7,48$$

und für die zweite Gruppe

$$a' = -9,6, \quad b' = -0,148.$$

Abbildung 5: Häufigkeitsverteilung der Summen von sieben Untertests des LPS in den Gruppen 1 und 2 (nach REINERT et al.).

Abbildung 5.1

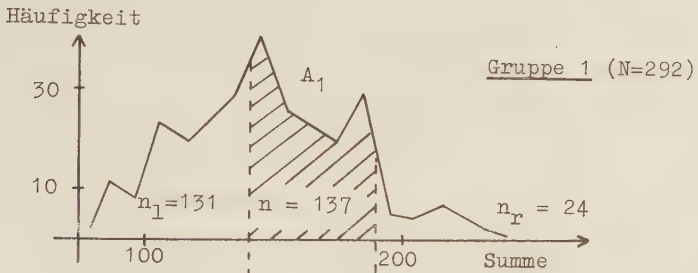
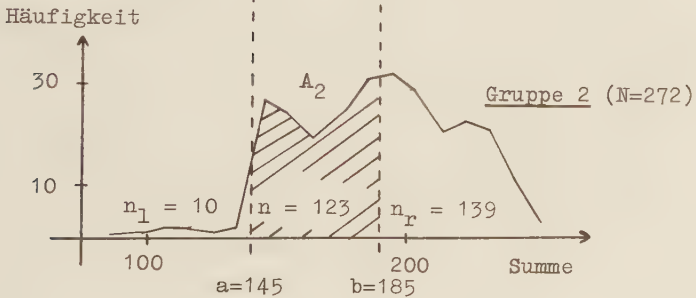


Abbildung 5.2



3. Schritt: Simulation der Selektion hinsichtlich der Gruppen A_1 und A_2 .

Im ersten Simulationsexperiment wurden zur Fingierung der Gruppe A_1 293 "Probanden" auf die im ersten Schritt beschriebene Weise erzeugt. Für jeden dieser Probanden wurde geprüft, ob die Summe x_0 seiner "Einzeltestwerte" die Bedingung

$$a' = -0,692 \leq x_0 \leq 7,48 = b'$$

erfüllte. Traf diese Bedingung zu, wurde der betreffende "Proband" der Auswahlstichprobe einverleibt, an Hand derer schließlich die empirischen Korrelationskoeffizienten zwischen den Variablen $X_1, \dots, X_7$ ermittelt wurden. Da 132 "Probanden" hinsichtlich ihrer x_0 -Werte links von a' und 24 rechts von b' zu liegen kamen, hatte diese Auswahlstichprobe wie bei REINERT et al. einen Umfang von 137.

Die Auswahlbedingungen bei der Gruppe A_2 wurden analog simuliert. Hier betrug die Zahl der insgesamt erzeugten Probanden 279. Drei Probanden überschritten hinsichtlich ihrer x_0 -Werte den zugelassenen Variationsbereich auf der linken und 153 auf der rechten Seite, so daß die Auswahlstichprobe einen Umfang von 123 hatte.

In Tabelle 4 sind die Ergebnisse beider Simulationsexperimente zu sehen. Die Übereinstimmung mit den von REINERT et al. mitgeteilten Korrelationen (s. Tabelle 2) ist der Tendenz nach vorhanden: Auch die bei den Simulationen gefundenen Korrelationskoeffizienten erweisen sich als betragsmäßig gering; vor allen Dingen aber findet man die erwarteten negativen Korrelationskoeffizienten wieder. Daß die zahlenmäßige Übereinstimmung einzelner einander entsprechender Korrelationskoeffizienten nicht noch höher ausgefallen ist, liegt vermutlich daran, daß REINERT et al. Jugendliche untersuchten, während die Ausgangswerte für die Simulationen an Erwachsenen gewonnen wurden. Ferner ist bei Stichprobenumfängen von 137 bzw. 123 noch mit beträchtlichen Stichprobenfehlern zu rechnen.

Tabelle 4: Simulierte Korrelationskoeffizienten.

Oberhalb der Diagonale stehen die Korrelationen, die den Auswahlbedingungen der Gruppe A_1 bei REINERT et al. entsprechen; darunter diejenigen der Gruppe A_2 .

	1	2	3	4	5	6	7
1		0,01	0,21	-0,12	0,36	-0,01	-0,03
2	-0,01		0,26	0,15	0,14	-0,13	0,03
3	0,36	0,16		0,06	0,28	-0,22	-0,22
4	-0,17	0,29	0,16		-0,10	0,14	-0,08
5	0,52	0,07	0,32	-0,10		-0,05	0,01
6	-0,08	0,02	-0,14	0,11	-0,05		0,21
7	0,14	0,03	-0,05	0,06	0,12	0,25	

Die Korrelationskoeffizienten in Tabelle 4 wurden noch einer Faktorenanalyse unterworfen, die nach Extraktion von drei gemeinsamen Faktoren abgebrochen wurde.

Die Matrizen der Ladungen auf diesen Faktoren wurden nach dem Varimaxkriterium rotiert und sind in den Tabellen 6 und 7 zu finden. Bei Betrachtung dieser Ladungsmatrizen fällt als erstes die Tendenz zumindest des zweiten Faktors zur Bipolarität auf. In der Ausgangsmatrix (s. Tabelle 5) haben sämtliche Ladungen gleiches Vorzeichen. Diese Tendenz zur Bipolarität ist auch bei den von REINERT et al. (1965, S. 280) rotierten Ladungsmatrizen vorhanden. Ferner sind die Kommunalitäten bei den Simulationen vergleichsweise gering, was auch von REINERT et al. (1965, S. 278 und S. 288) hinsichtlich ihrer Untersuchungsergebnisse erwähnt wird.

Tabelle 5: Faktorenanalyse der originalen Korrelationen zwischen sieben Untertests des LPS (s. Tabelle 3):

Matrix A_{73} der Ladungen (varimax-rotiert) auf drei gemeinsamen Faktoren.

	Faktor			h^2
	G_1	G_2	G_3	
1	0,75	0,35	0,24	0,74
2	0,41	0,30	0,62	0,64
3	0,60	0,20	0,46	0,61
4	0,22	0,30	0,71	0,63
5	0,77	0,35	0,28	0,79
6	0,24	0,60	0,30	0,51
7	0,34	0,57	0,30	0,50
Durchschnitt				0,63

Tabelle 6.: Faktorenanalyse der Korrelationskoeffizienten, erhalten bei der Simulation der Gruppe A_1 :

Matrix der Ladungen (varimax-rotiert) auf drei gemeinsamen Faktoren.

	Faktor			h^2
	G_1	G_2	G_3	
1	0,56	0,00	-0,08	0,33
2	0,14	-0,12	0,44	0,23
3	0,39	-0,42	0,37	0,47
4	-0,15	0,03	0,38	0,17
5	0,67	-0,00	0,06	0,45
6	-0,04	0,69	0,07	0,48
7	0,03	0,34	-0,09	0,13
Durchschnitt				0,32

Tabelle 7: Faktorenanalyse der Korrelationskoeffizienten, erhalten bei der Simulation der Gruppe A₂:
Matrix der Ladungen (varimax-rotiert) auf drei gemeinsamen Faktoren.

	Faktor			h ²
	G ₁	G ₂	G ₃	
1	0,74	0,05	-0,13	0,56
2	0,06	0,05	0,50	0,26
3	0,50	-0,24	0,34	0,42
4	-0,12	0,06	0,60	0,38
5	0,71	0,06	-0,02	0,51
6	-0,10	0,48	0,06	0,25
7	0,13	0,52	0,03	0,29
Durchschnitt				0,38

3. Beispiel: Die "Erzeugung" von Faktoren durch Selektion.

Das dritte Beispiel soll zeigen, wie weit durch bestimmte Selektionen bzw. Vorgänge, die Selektionen analog sind, das Ergebnis einer Faktorenanalyse bestimmt wird. Zu diesem Zweck seien 10 Zufallsvariable $X_1, X_2, \dots, X_{10}$, die man als Intelligenztests interpretieren möge, zugrundegelegt. Entsprechend der SPEARMANschen Zweifaktorentheorie der Intelligenz (s. S. 1) sei angenommen, daß diese Variablen sich wie folgt zusammensetzen:

$$(a) \quad X_i = 0,8 G_1 + 0,6 H_i, \quad i = 1, 2, \dots, 10.$$

Hierbei bedeuten G_1 den Generalfaktor und $H_1, H_2, \dots, H_{10}$ die testspezifischen Faktoren. Sämtliche Faktoren seien voneinander unabhängig und mögen die Varianz 1 besitzen. Somit sind die Kommunalitäten der $X_1, X_2, \dots, X_{10}$ und die Korrelationskoeffizienten zwischen ihnen sämtlich gleich 0,64.

Zwei Selektionsvariable seien wie folgt definiert:

$$Y_1 =: H_1 + H_2 + H_3 + H_4 + H_5$$

$$Y_2 =: H_6 + H_7 + H_8 + H_9 + H_{10}$$

Beide Selektionsvariable besitzen ursprünglich die Varianz 5. Infolge von Selektion möge die Varianz von Y_1 auf das 3-fache vergrößert, die Varianz von Y_2 dagegen auf ein Viertel verringert werden. Wie das im einzelnen geschieht, kann hier offenbleiben. Es sei lediglich bemerkt, daß unter bestimmten Bedingungen, z. B. durch Lernen und Übung, Varianzänderungen dieser Art auftreten können.

Die (13-dimensionale) Variable ($Y_1, Y_2, G_1, H_1, \dots, H_{10}$) besitzt die in Tabelle 8 wiedergegebene Kovarianzmatrix.

Tabelle 8: Kovarianzmatrix der zufälligen 13-dimensionalen Variable ($Y_1, Y_2, G_1, H_1, \dots, H_{10}$)

	Y_1	Y_2	G_1	H_1	...	H_5	H_6	...	H_{10}
Y_1	5	0	0	1	...	1	0	...	0
Y_2	0	5	0	0	...	0	1	...	1
G_1	0	0	1	0	...	0	0	...	0
H_1	1	0	0	1	...	0	0	...	0
.	.	.	.	.	...	.	.	...	.
.	.	.	.	.	...	.	.	...	.
.	.	.	.	.	...	.	.	...	.
H_5	1	0	0	0	...	1	0	...	0
H_6	0	1	0	0	...	0	1	...	0
.	.	.	.	.	...	.	.	...	.
.	.	.	.	.	...	.	.	...	.
.	.	.	.	.	...	.	.	...	.
H_{10}	0	1	0	0	...	0	0	...	1

Auf diese Matrix können die LAWLEYSchen Selektionsformeln (19) (s. S. 82) angewandt werden. Hierbei entspricht die linke obere Teilmatrix der Matrix V_{pp} , die infolge der

Selektion in

$$V_{pp}^* = \begin{pmatrix} 15 & 0 \\ 0 & 1,25 \end{pmatrix}$$

übergeht. Man erhält so die "neue" Kovarianzmatrix der Faktorvariable ($G_1, H_1, \dots, H_{10}$), mit deren Hilfe man zusammen mit (a) gem. (4.1) die Kovarianzmatrix von ($X_1, X_2, \dots, X_{10}$) nach Selektion berechnen kann.

Nach Subtraktion der "neuen" spezifischen Varianzanteile von der Varianz der Testvariablen und Transformation der Kovarianzmatrix in die Korrelationsmatrix gem. (13.2) erhält man die reduzierte Korrelationsmatrix $\bar{R}$ von ($X_1, X_2, \dots, X_{10}$). Sie ist in Tabelle 9 zu finden.

Tabelle 9: Reduzierte Korrelationsmatrix von ($X_1, X_2, \dots, X_{10}$) nach Selektion.

	X_1	X_2	...	X_5	X_6	...	X_{10}
X_1	0,56	0,69	...	0,69	0,62	...	0,62
X_2	0,69	0,56	...	0,69	0,62	...	0,62
.	.	.		.	.		.
.	.	.		.	.		.
.	.	.		.	.		.
X_5	0,69	0,69	...	0,56	0,62	...	0,62
X_6	0,62	0,62	...	0,62	0,67	...	0,62
.	.	.		.	.		.
.	.	.		.	.		.
.	.	.		.	.		.
X_{10}	0,62	0,62	...	0,62	0,62	...	0,67

Nach der Hauptachsentransformation dieser Matrix erhält man als Eigenwerte:

6,3 0,15 0,06 0,06 0,06 0,06 -0.13 -0.13 -0.13 -0.13

Da alle Eigenwerte verschieden von Null sind, ist der Rang von $\bar{R}$ gleich 10, obwohl nur ein gemeinsamer Faktor ursprünglich unterlegt worden ist. Ein solches Ergebnis war nach (7.4) zu erwarten. Ferner erweist sich $\bar{R}$ nicht als positiv semi-definit, da negative Eigenwerte vorhanden sind. Es ist daher zweckmäßig, die vollständige Korrelationsmatrix R , die mit Sicherheit positiv semi-definit ist, einer Faktorenanalyse zu unterwerfen. Nach Einsetzen von Einsen in die Hauptdiagonale von $\bar{R}$ erhält man mit Hilfe der Hauptachsenmethode ein Faktorenmuster, von dem in Tabelle 10 nur die ersten zwei Spalten mitgeteilt sind. Die übrigen Faktoren liefern keinen nennenswerten Beitrag zur Gesamtvarianz. (Eigenwerte kleiner als 0,5.) Es fällt die Bipolarität des zweiten Faktors auf, die die durchgeführte Selektion deutlich widerspiegelt.

Tabelle 10: Faktorenanalyse der Korrelationskoeffizienten aus Tabelle 9.

Variable	Faktor		h^2
	F_1	F_2	
X_1	0,84	-0,23	0,76
.	.	.	.
.	.	.	.
.	.	.	.
X_5	0,84	-0,23	0,76
X_6	0,80	0,24	0,70
.	.	.	.
.	.	.	.
.	.	.	.
X_{10}	0,80	0,24	0,70
Eigenwerte	6,67	0,53	

Zum Vergleich dazu wurde eine Korrelationsmatrix von 10 der 11 Untertests des Hamburg-Wechsler-Intelligenztests für Erwachsene (HAWIE), die im Handbuch zum HAWIE (WECHSLER,

1956, S. 214) mitgeteilt wird, einer Faktorenanalyse unterworfen. Bei den Untertests handelt es sich im einzelnen um folgende:

Allgemeines Wissen (AW), Allgemeines Verständnis (AV), Zahlennachsprechen (ZN), Rechnerisches Denken (RD), Gemeinsamkeitenfinden (GE), Bilderordnen (BO), Bilderergänzen (BE), Mosaik-Test (MT), Figurenlegen (FL) und Zahlen-Symbol-Test (ZS). Die Stichprobe bestand aus 370 Erwachsenen im Alter von 20 bis 34 Jahren. Unter Anwendung desselben Abbruchkriteriums wie oben (Zahl der Faktoren gleich Zahl der Eigenwerte größer als 0,5) ergab die Analyse ebenfalls zwei Faktoren (F_1 , F_2). Die Ladungen der Tests auf diesen Faktoren sind in Tabelle 11 aufgeführt. Auch hier ist der zweite Faktor bipolar.

Tabelle 11: Faktorenanalyse der von WECHSLER (1956) mitgeteilten Korrelationskoeffizienten von 10 Untertests des HAWIE.

	Faktor		h^2
	I	II	
AW	0,84	-0,30	0,80
AV	0,76	-0,19	0,61
ZN	0,60	-0,10	0,37
RD	0,80	-0,14	0,66
GE	0,85	-0,22	0,77
BO	0,65	0,18	0,45
BE	0,75	0,22	0,61
MT	0,78	0,29	0,69
FL	0,59	0,36	0,48
ZS	0,68	0,01	0,46
Eigenwerte	5,41	0,50	

In diesem Zusammenhang ist es bemerkenswert, daß die ersten fünf Untertests beim HAWIE den sogenannten Verbalteil und die letzten fünf den sogenannten Handlungsteil ausmachen.

Literaturverzeichnis

- AITKEN, A. C. Note on selection from a multivariate normal population. Proc. Edin. Math. Soc., 1935, 4, 106-110.
- ALBERT, A. A. The matrices of factor analysis. Proc. nat. acad. Sci., 1944, 30, 90-95.
- ALBERT, A. A. The minimum rank of a correlation matrix. Proc. nat. acad. Sci., 1944, 30, 144-146.
- AHMAVAARA, Y. The mathematical theory of factorial invariance under selection. Psychometrika, 1954, 19, 27-38.
- BIRNBAUM, Z. W. Effects of linear truncation on a multinormal population. Ann. math. Stat., 1950, 21, 272-279.
- BIRNBAUM, Z. W., E. PAULSON & F. C. ANDREWS. On the effect of selection performed on some coordinates of a multidimensional population. Psychometrika, 1950, 15, 191-204.
- BORGATTA, E. F., L. S. COTTRELL & H. J. MEYER. On the dimensions of group behaviour. Sociometry, 1956, 19, 223-240.
- BURT, C. The relations of educational abilities. Brit. J. ed. Psychol., 1939, 2, 45-71.
- BURT, C. The factors of mind. London: Univ. of London Press, 1940.

- BURT, C. & W. L. GREGORY. Scientific methods in Psychology II. Brit. J. stat. Psychol., 1958, 11, 105-128.
- BURT, C. The stability of factors. Brit. J. stat. Psychol., 1964, 17, 178-180.
- CAMPBELL, F. L. A study of truncated bivariate normal distributions. (Doctoral dissertation) University of Michigan, 1945 (zit. nach COHEN, 1957).
- CATTELL, R. B. Parallel-proportional profiles and other principles for determining the choice of factors by rotation. Psychometrika, 1944, 9, 267-283.
- CATTELL, R. B. The dimension of culture patterns by factorization of national characters. J. abn. soc. Psychol., 1949, 44, 443-469.
- CATTELL, R. B. A note on factor-invariance. Brit. J. stat. Psychol., 1949, 2, 134-139.
- CATTELL, R. B. & A. K. S. CATTELL. Factor solutions for proportional profiles: Analytical solution and an example. Brit. J. stat. Psychol., 1955, 8, 83-92.
- CATTELL, R. B. & R. L. GORSUCH. The definition and measurement of national morale and morality. J. soc. Psychol., 1965, 67, 77-96.
- COHEN, A. C. Restriction and selection in multinormal distributions. Ann. math. stat., 1957, 28, 731-741.
- COHEN, A. C. On the solution of estimating equations for truncated and censored samples from normal populations. Biometrika, 1957, 44, 225-236.

- COMREY, A., L. M. R. MESCHIERI & R. NENCINI. A comparison of personality factor structure in American and Italian subjects. *J. pers. Soc. Psychol.*, 1965, 1, 257-261.
- DUNN, J. A. Stability of the factor structure of the test anxiety scale for children across age and sex groups. *J. consult. Psychol.*, 1965, 29, 187.
- ELMORE, T. M. The development of a scale to measure psychological anomie and its implications for counseling psychology. *Pro. 73. Ann. Conv. Am. Psychol. Assoc.*, 1965, 359-360.
- EYSENCK, H. J. The logical basis of factor analysis. *Am. Psychol.*, 1953, 8, 105-114.
- FAHRENBERG, J. & M. MYRTEK. Ein kritischer Beitrag zur psychophysiologischen Persönlichkeitsforschung. *Z. exp. angew. Psychol.*, 1966, 13, 222-247.
- FISCHER, G. H. Zum Problem der Interpretation faktorenanalytischer Ergebnisse. *Psychol. Beitr.*, 1967, 10, 133-135.
- FISCHER, G. & J. ROPPERT. Monte-Carlo-Untersuchungen an einem Intelligenzmodell mit vier Faktoren. *Ber. 24. Kongr. DGfPs, Wien. Göttingen: Hogrefe*, 1964.
- FLEISHMAN, E. A. Dimensional Analysis of psychomotor abilities. *J. exp. Psychol.*, 1954, 48, 437-454.
- GANTMACHER, F. R. *Matrizenrechnung I*. Berlin: VEB Deutscher Verlag der Wissenschaften, 1965.
- GUILFORD, F. P. The structure of intellect. *Psychol. Bull.*, 1956, 53, 267-293.

- HARMAN, H. H. The square root method and multiple groups methods of factor analysis. *Psychometrika*, 1954, 19, 39-55.
- HARMAN, H. H. Modern factor analysis. Chicago: Univer. of Chicago Press, 1960.
- HARMAN, H. H. Factor analysis by minimizing residuals (minres). *Psychometrika*, 1966, 31, 351-368.
- HEDINGER, U. K. Die Faktorenstruktur komplexer Denkaufgaben. *Z. exp. angew. Psychol.*, 1965, 12, 337-403.
- HOEPFNER, R., K. NIHIRA & J. P. GUILFORD. Intellectual abilities of symbolic and semantic judgement. *Psychological Monographs*, 1966, 80, No. 16.
- HORN, W. Leistungsprüfsystem. L-P-S. Handanweisung für die Durchführung, Auswertung und Interpretation. Göttingen: Hogrefe, 1962.
- HORN, W. Begabungstests system. B-T-S. Handanweisung, Durchführung, Auswertung und Interpretation. Göttingen: Hogrefe, 1965.
- HOTELLING, H. Analysis of a complex of statistical variables into principal components. *J. exp. Psychol.*, 1933, 24, 417-441 und 498-520.
- HÜRSCH, L. Der Einfluß nicht-kognitiver Versuchsbedingungen auf die Faktorenstruktur von Intelligenzleistungen. *Schweiz. Z. Psychol.*, 1967, 26, 61-66.
- KALLINA, H. Das Unbehagen in der Faktorenanalyse. *Psychol. Beitr.*, 1967, 10, 81-86.

- KALVERAM, K. Th. Über die Beurteilung von Persönlichkeitseigenschaften nach dem Paarvergleich. Unveröff. Staatsexamensarbeit, Marburg/L., 1962.
- KALVERAM, K. Th. Veränderungen von Faktorenstrukturen durch "simultane" Überlagerung. Arch. ges. Psychol., 1965, 117, 296-305.
- KRISTOF, W. Die Beziehungen zwischen mehrdimensionaler Skalierung und Faktorenanalyse. Psychol. Beitr., 1963, 7, 370-378.
- KRISTOF, W. Die faktorielle Struktur geometrisch optischer Täuschungen. Z. exp. angew. Psychol., 1964, 11, 71-94.
- KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 1964, 29, 1-27.
- LAWLEY, D. N. The estimation of factor loadings by the method of maximum likelihood. Proc. Roy. Soc. Edin. A, 1940, 60, 64-82.
- LAWLEY, D. N. A note on Karl Pearson's selection formulae. Proc. Roy. Soc. Edin. A, 1943, 62, 28-30.
- LAWLEY, D. N. Estimation in factor analysis under various initial assumptions. Brit. J. stat. Psychol., 1958, 11, 1-12.
- LEDERMANN, W. On the rank of the reduced correlation matrix in multiple factor analysis. Psychometrika, 1937, 2, 85-93.
- LEVIN, J. Three-mode factor analysis. Psychol. Bull., 1965, 64, 442-452.

- LIENERT, G. A. Die Faktorenstruktur der Intelligenz als Funktion des Neurotizismus. Z. exp. angew. Psychol., 1963, 10, 140-159.
- LIENERT, G. A. Belastung und Regression. Meisenheim: Hain, 1964.
- MEREDITH, W. Notes on factorial invariance. Psychometrika, 1964, 29, 177-185.
- MEREDITH, W. A method for studying differences between groups. Psychometrika, 1965, 30, 15-29.
- MERZ, F. Über die soziale Determination von Eigenschaften. Jahrb. Psychol. Psychotherapie, 1957, 2, 183-196.
- ORLIK, P. Das Dilemma in der Faktorenanalyse - Zeichen einer Aufbaukrise in der modernen Psychologie. Psychol. Beitr., 1967, 10, 87-98.
- PEARSON, K. On the influence of natural selection on the variability and correlation of organs. Phil. trans. Roy. Soc. A, 1903, 200, 1-66.
- PEARSON, K. The general theory of the influence of selection on correlation and variation. Biometrika, 1912, 8, 437-443.
- RAO, C. R. Estimation and tests of significance in factor analysis. Psychometrika, 1955, 20, 93-111.
- RAO, C. R. Some statistical methods for comparison of growth curves. Biometrics, 1958, 3, 37-40.
- REINERT, G., P. B. BALTES & L. R. SCHMIDT. Faktorenanalytische Untersuchungen zur Differenzierungshypothese der Intelligenz: Die Leistungsdifferenzierungshypothese. Psychol. Forsch., 1965, 28, 246-300.

- SCHMETTERER, L. Einführung in die mathematische Statistik.
Wien: Springer-Verlag, 1956.
- SCHUESSLER, K. F. & H. DRIVER. A factor analysis of primitive societies. Amer. sociological Rev., 1956, 21, 493-499.
- SEIFERT, Th. Faktorenanalyse einiger Schriftmerkmale. Z. exp. angew. Psychol., 1964, 11, 645-666.
- SELG, H. & G. LISCHKE. Eine Faktorenanalyse von Aggressionsvariablen. Z. exp. angew. Psychol., 1966, 15, 506.
- SPEARMAN, C. General intelligence, objectively determined and measured. Amer. J. Psychol., 1904, 15, 201-293.
- TAYLOR, R. G. & W. FARQUHAR. Personality, motivation and achievement: Theoretical constructs and empirical factors. J. counsel. Psychol., 1965, 12, 186-191.
- THOMSON, G. H. & W. LEDERMANN. The influence of multivariate selection on the factorial analysis of ability. Brit. J. Psychol., 1939, 29, 288-305.
- THURSTONE, L. L. Multiple factor analysis. Psychol. Rev., 1931, 38, 406-427.
- THURSTONE, L. L. The effect of selection in factor analysis. Psychometrika, 1945, 10, 165-198.
- THURSTONE, L. L. Multiple factor analysis. Chicago: Univer. of Chicago Press, 1947.
- TUCKER, L. R. Determination of parameters of a functional relation by factor analysis. Psychometrika, 1958, 23, 19-23.

- TUCKER, L. R. Determination of generalized learning curves by factor analysis. Princeton (Princeton Univer. Test. Serv.), 1960.
- TUCKER, L. R. The extension of factor analysis to three-dimensional matrices. In: N. FREDERIKSEN & H. GULLIKSEN (Eds.): Contribution to mathematical Psychology. New York: Holt, Rinehart & Winston, 1964, 109-127.
- VOTAW, D. F., F. A. RAFFERTY & W. L. DEEMER. Estimation of parameters in a truncated trivariate normal distribution. Psychometrika, 1950, 15, 339-347.
- VUKOVICH, A. Untersuchungen zur Invarianz der Faktorenstruktur von Begabungstests bei systematischer Variation der Probandengruppen. Ber. 24. Kongr. DGfPs. in Wien, 1964. Göttingen: Hogrefe, 1965, 173-177.
- VUKOVICH, A. Faktorielle Typenbestimmung. Psychol. Beitr., 1967, 10, 112-121.
- WECHSLER, D. Die Messung der Intelligenz Erwachsener (bearb. von HARDESTY, Anne & H. LAUBER). Bern und Stuttgart: Huber, 1956.
- WHERRY, R. J., J. C. NAYLOR, R. J. (Jr.) WHERRY & R. F. FALLIS. Generating multiple samples of multivariate data with arbitrary population parameters. Psychometrika, 1965, 30, 303-313.
- WRIGLEY, C. The distinction between common and specific variance in factor theory. Brit. J. stat. Psychol., 1957, 10, 81-98.

L e b e n s l a u f .

Als viertes Kind der Eheleute Heinrich Kalveram und Wilhelmine, geb. Heisterkamp, wurde ich, Karl Theo Kalveram, am 12. Februar 1935 in Duisburg-Hamborn geboren.

Von Herbst 1941 bis Ostern 1946 besuchte ich die Volksschulen in Hamborn, Bürgeln bei Marburg, Sigma-
ringen und Laucherthal. Ostern 1946 wurde ich in die Sexta der jetzigen Leibnizschule, eines naturwissen-
schaftlichen Gymnasiums in Hamborn, aufgenommen. An
dieser Schule bestand ich 1954 die Reifeprüfung.

Mit dem folgenden Sommersemester 1954 begann ich,
Mathematik und Physik zu studieren, und zwar die ersten
beiden Semester in Bonn, die folgenden in Marburg.

Nachdem ich im Sommer 1963 das Staatsexamen ein-
schließlich der Ergänzungsprüfung in Psychologie be-
standen hatte, begann ich mit dem regulären Studium
der Psychologie in Marburg. Hier legte ich im Dezem-
ber 1965 auch die Diplom-Hauptprüfung für Psychologen
ab. Seither verwalte ich eine wissenschaftliche Assi-
stentenstelle am Marburger Institut für Psychologie.

Marburg, den 27. Dezember 1967

Karl Theo Kalveram

